INFOSECTRAIN
Educate. Excel. Empower.

ISACA AAISM

STUDY GUIDE

LAST MINUTE EXAM PREPARATION



DOMAIN 1	DOMAIN 2	DOMAIN 3
AI Governance & Program Management	AI Risk & Opportunity Management	AI Technologies & Controls
31% of exam	**31%** of exam	**38%** of exam

START HERE

How to Use This Guide

This guide is built for the final **3–7 days** before your exam. It is not a textbook: every page is compressed to the ideas the exam rewards — judgement, priorities and the “manager’s answer”.

Your revision plan

When	What to do
Day 1–3	Read one domain per day end-to-end. Mark every **Exam Trap** you got wrong.
Day 4–5	Re-read only the boxes and tables. Redraw the diagrams from memory.
Night before	Read the three one-page **Cram Sheets** and the **Final-Hour Checklist**; nothing else.

Box legend

KEY CONCEPT
Definitions and models you must know.

EXAM TRAP
Attractive wrong answers and the reason they fail.

MANAGER MINDSET
How an AI security manager would decide.

REMEMBER
Short memory hooks.

Exam snapshot

Item	Detail
Questions	~90 multiple-choice, single best answer
Time	150 minutes (≈ 1.6 minutes per question)
Scoring	Scaled 200–800; **450 to pass**
Eligibility	Active CISM or CISSP
Domain 1	AI Governance & Program Management — **31%**
Domain 2	AI Risk (& Opportunity) Management — **31%**
Domain 3	AI Technologies & Controls — **38%**

Verify exam logistics on isaca.org before booking. ISACA can update the outline.

NAVIGATE

Contents

Think Like an AI Security Manager 4

The Seven Decision Rules ... 4
Decoding the Question Stem ... 4
The Golden Thread ... 5

AI Governance & Program Management 6

A · Stakeholders, Frameworks & Regulatory Requirements ... 6
B · AI Strategy, Policies & Responsible AI ... 13
C · AI Asset & Data Life Cycle Management ... 17
D · AI Security Program Development & Management ... 19
E · Business Continuity & AI Incident Response ... 21
Domain 1 Cram Sheet: AI Governance (31%) ... 24

AI Risk & Opportunity Management 25

A · Risk Frameworks, Acceptable Limits & Responses ... 25
B · Threat & Vulnerability Management ... 33
C · AI Vendor & Supply Chain Management ... 36
Domain 2 Cram Sheet: AI Risk Management (31%) ... 40

AI Technologies & Controls 41

A · AI Architecture, Models & Secure Design ... 41
B · The AI Life Cycle ... 45
C · Data Management Controls ... 48
D · Privacy, Ethics, Trust & Safety Controls ... 49
E · Security Controls & Monitoring ... 51
Domain 3 Cram Sheet: AI Technologies & Controls (38%) ... 56

Glossary of High-Yield Terms 57

Final-Hour Checklist 59

MINDSET

Think Like an AI Security Manager

AAISM sits on top of CISM/CISSP. The exam assumes you already know security; it tests whether you can **govern AI**. Most wrong answers are technically true but are the wrong level, the wrong order or the wrong owner.

The Seven Decision Rules

#	Rule	What it means
1	Governance before technology	Strategy, policy, ownership and risk appetite come before tools.
2	Visibility before control	You cannot protect an AI system you have not inventoried. Inventory, then assess, then treat.
3	Business context drives risk	The same model is low-risk in marketing and high-risk in hiring or lending.
4	Accountability never transfers	Buying AI, outsourcing it or ensuring it does not move accountability off the organisation that deploys it.
5	People first	When human safety or fundamental rights are at stake, protecting people outranks availability, cost and speed.
6	Prevent and prepare	Pre-deployment assessment, testing and documented thresholds beat after-the-fact fixes.
7	Continuous, not one-off	AI drifts. Monitoring, KRIs and periodic review are part of every correct control set.

Decoding the Question Stem

Keyword	What it is really asking
FIRST	The earliest logical step; usually **understand, identify, inventory or assess** before acting.
BEST / MOST appropriate	The option that solves the **root cause** at the right level, with lasting effect.
PRIMARY / MAIN purpose	The reason the control exists, not a side benefit.
GREATEST risk / concern	Highest impact to **people, compliance or the business**; not the most technical.
MOST effective control	Preventive > detective > corrective, if they address the same cause.
Who is accountable	Board / senior management for risk appetite; business owner for the use case; data owner for data.

EXAM TRAP: CLASSIC DISTRACTOR PATTERNS

- "Deploy a tool" when the scenario lacks a policy, owner or inventory.
- "Shut it down immediately" when proportionate mitigation is available and people are not at risk.
- "Check firewall logs" for a biased or drifting model; that is a data or model problem.
- "Rely on the vendor's certification" in place of your own due diligence.
- Answers that are true but sit at the technician level, not the manager level.

The Golden Thread



Most "what should happen FIRST/NEXT" questions follow this order.

MANAGER MINDSET: THE ONE-LINE TEST

Before choosing, ask: **"Would a CISO say this to the board?"** If an option is a configuration step, a single technical tweak, or skips approval and ownership, it is rarely the BEST answer for a governance-level question. In Domain 3, technical controls are fine, but still pick the one that addresses the root cause and fits the risk.

DOMAIN 1

AI Governance & Program Management

31% OF EXAM

Stakeholders · frameworks & laws · strategy & policy · responsible AI · asset & data life cycle · security program · AI incident response

A Stakeholders, Frameworks & Regulatory Requirements

Why AI needs its own governance

AI governance is the senior-level oversight, processes, standards and safeguards that keep AI use safe, ethical, lawful and value-creating. It exists because employees will adopt AI tools whether or not a policy exists, so the governing body must set direction, boundaries and accountability *before* unmanaged use becomes the norm. Governance goes beyond legal compliance: it also covers model drift, hallucinations, long-term reliability and social responsibility.

WHAT MAKES AI DIFFERENT FROM ORDINARY IT

- Behaviour is **learned from data**, not coded. Data quality becomes a control.
- Outputs are **probabilistic**; errors are expected and must be thresholded.
- Models change over time (drift, retraining) faster than compliance processes.
- Decisions can be opaque; **explainability** is a governance requirement.
- Impact reaches people and society (bias, rights, jobs, environment).

ENTERPRISE CONSIDERATIONS RAISED BY AI

- **Data confidence:** accuracy, completeness, integrity of inputs.
- **Transparency** of decisions that replace human judgement (e.g., credit).
- Competitive pressure to adopt vs. AI decision risk.
- Governance/compliance gap: AI evolves faster than controls.
- Workforce impact; brand and reputation; a **shifted risk profile** (new risks appear, some old ones such as repetitive human error shrink).

AI readiness: the first step before adoption

Before scaling AI, the governing body asks structured readiness questions; management analyses the answers, identifies gaps and launches corrective action. Readiness has three lenses:

Lens	Typical questions
Mission & value	How does AI support the vision and strategy? What need does it address? How are benefits and ROI measured, realised and tracked? Is the culture ready (talent, change management)?
Risk, compliance & ethics	Is AI risk in the enterprise risk register and consistent with risk appetite? Are fairness, transparency and accountability frameworks in place? Are we tracking AI regulation and contractual impact? Is sensitive data protected? Has a governance framework been formally adopted?
Resources & enablement	Are people, time and budget allocated? Are skills, platforms and guardrails available? Is HR aligned? Are project planning, reporting and KPIs defined?

KEY CONCEPT: FOUR FOUNDATIONS OF A DEFINED AI GOVERNANCE SYSTEM

1. Mature **I&T governance** (stakeholder needs → balanced objectives → direction → monitoring).
2. Mature **management processes** (innovation, portfolio, external compliance).
3. Established **structures** (board, governance, strategy, risk and audit committees).
4. **Skilled resources** able to make informed AI decisions.

Roles and responsibilities

Category	Roles	What they own
Leadership & strategy	Executive management · Chief AI Officer (CAIO) · AI steering committee	Tone at the top; approve AI strategy; ultimate accountability for AI impact. The CAIO leads adoption (standalone or combined role). The steering committee is cross-functional and may sit inside an IT steering or innovation committee.
Development & operations	IT development (data scientists, engineers) · IT operations · product management	Build, deploy and maintain AI; infrastructure; product managers bridge technical teams and users.
Users	End users · HR · customer service	Use AI; HR handles job classification, skills and training; customer service supports external users.
Governance & oversight	Governance committee · risk management · information security & privacy · internal audit	Approve/monitor policy and metrics; integrate AI into ERM; apply security/privacy controls across the life cycle; independent assurance.

KEY CONCEPT: THE GOVERNING BODY

The board sets financial and non-financial objectives (culture, values, ethics) and is **fully accountable** for all organisational activity; this accountability **cannot be delegated**. For AI it must:

- confirm intended use fits risk appetite;
- be ready to **modify, suspend or abort** AI initiatives as risk evolves;
- stay informed; and
- avoid **anthropomorphising** AI (attributing human judgement or morality to it).

Oversight duties: management oversight · supervision · evaluation · reporting · change-management oversight.

Stakeholders

Successful AI needs stakeholder **mapping** → **prioritisation** (risk assessment, gap analysis) → **engagement** → **ethical development** → **monitoring**.

Stakeholder	Primary concern
Board / senior leadership	Strategic alignment, value (cost, efficiency, growth), implementation and reputational risk, compliance, workforce fears.
Employees	Training/upskilling; AI that complements rather than disrupts workflows.
IT & technical teams	Secure infrastructure, alignment with business/technology/security strategy, data readiness (quality, quantity, access), long-term support.
Customers / end users	Usability, privacy (e.g., GDPR, CPRA), transparency about how AI shapes decisions, fairness.
Third parties	Integration, shared goals, ethical AI practices aligned to your values.
Regulators & policymakers	Compliance, proactive engagement, transparent reporting of AI impacts.
Society & communities	Social responsibility, inclusion of marginalised groups, environmental footprint, misuse.

AI charter and AI steering committee

■ AI CHARTER: THE "WHY, WHAT, WHO"

A formal document that translates governance intent into scope, objectives and accountability.

1. Project name & description (problem being solved)
2. Objectives: specific, measurable, achievable
3. Scope: in-scope and out-of-scope (prevents scope creep)
4. Stakeholders (internal + external)
5. Governance structure: roles, reporting, escalation, decision rights, change accountability
6. Timeline & milestones
7. Resources & budget
8. Risk management
9. Success metrics (ROI, accuracy, adoption, KPIs)
10. Approval & sign-off by executives/sponsors

■ AI STEERING COMMITTEE

Purpose: strategic oversight during and after the project life cycle.

Decides: budget changes, scope adjustments, major risk decisions, resolution of significant issues; approves policies on AI use, data handling and bias mitigation.

Members: chair (CIO/CTO/CDO), sponsors, business domain experts, technical experts (data science, security, architects), compliance/legal/risk/ethics.

Cadence: regular (monthly/quarterly) plus ad hoc for urgent decisions.

> **REMEMBER**
>
> **Charter** = the document. **Steering committee** = the body that steers it.

Who does what: a sample AI governance RACI

Activity	Board	Exec / CAIO	Steering cttee	Business owner	Security / privacy	Risk	Internal audit
Set AI risk appetite	A	R	C	C	C	C	I
Approve AI strategy	A	R	C	C	C	C	I
Approve AI policy / AUP	I	A	R	C	C	C	I
Approve a use case	I	A	R	R	C	C	I
Maintain AI inventory	—	A	I	R	C	C	I
Risk-assess an AI system	—	I	I	A	R	R	I
Apply security / privacy controls	—	I	I	A	R	C	I
Independent assurance	I	I	I	C	C	C	A/R

R = responsible, A = accountable, C = consulted, I = informed. Key point: the business owner is accountable for a use case's risk; security implements controls; audit stays independent.

> **KEY CONCEPT: PROVIDER VS. DEPLOYER**
>
> An enterprise can be an AI **provider** (develops a model/system and places it on the market, free or paid; owns design, development and compliance of the product) and/or an AI **deployer** (uses an AI system professionally; owns how it is implemented, how outputs are used and whether use is compliant and ethical). Responsibilities and risk exposure differ by role; many enterprises are both.

Standards and frameworks you must recognise

Reference	Type	One-line purpose
ISO/IEC 42001:2023	Certifiable management-system standard	Requirements for an AI Management System (AIMS): establish, implement, maintain and continually improve AI governance.
ISO/IEC 23894:2023	Guidance	AI risk management: integrates AI risk into ERM (builds on ISO 31000).
ISO/IEC 38507:2022	Guidance	Governance implications of AI use for the governing body.
ISO/IEC 22989:2022	Vocabulary	AI concepts and terminology.
ISO/IEC 23053:2022	Framework	Describes AI systems that use machine learning: common components and terms.
ISO/IEC TR 24027 / TR 24028 / 24029-1	Technical reports	Bias in AI · trustworthiness · robustness of neural networks.
IEEE 7000-2021	Process standard	Embedding ethical values into system design (value-based engineering).
OECD AI Principles	Intergovernmental policy	Five values-based principles + five recommendations to governments.
NIST AI RMF 1.0	Voluntary framework	Govern > Map > Measure > Manage for trustworthy AI; for developers and deployers.
Singapore Model AI Governance Framework (PDPC)	Practical guidance	Deployment-focused governance: accountability, transparency, human involvement.
COBIT	Governance framework	Holistic I&T governance; map AI onto EDM/APO/BAI/DSS/MEA rather than build a parallel structure.
CSA Four Pillars	Practitioner framework	Model cards · data sheets · risk cards · scenario planning, linked by a feedback loop.

EXAM TRAP

ISO/IEC 42001 is the certifiable AI management system; **23894** is risk guidance; **NIST AI RMF** is voluntary and not certifiable. Don't mix them up.

COBIT ACROSS THE AI LIFE CYCLE

Phase	COBIT domain
Design	EDM: align with strategy, direct resources, ethics
Develop	APO: structures, data/security policies (build or buy)
Deploy	BAI: secure implementation, testing, change (BAI06)
Operate	DSS: service levels, security (DSS05), continuity, incidents
Monitor	MEA: performance, compliance, improvement

Ethics: EDM01 (governance framework), EDM02 (benefit delivery incl. ethical impact), APO01 (management framework). **Assets:** BAI09. **Data/privacy:** APO14.

CSA FOUR PILLARS

- **Model cards:** objectives, training data, performance, limitations, adversarial resistance.
- **Data sheets:** dataset origin, creation, types, intended use, biases.
- **Risk cards:** observed risks, remediation, expected user behaviour.
- **Scenario planning:** explore misuse/failure before it happens.

Laws and regulation

AI-specific laws are new, largely untested in court and often sub-national. Non-compliance can hit developers *and* deployers/users. Organisations must track emerging rules, assess obligations and adapt controls. The main compliance themes for GenAI are what users put into the system:

Theme	Governance response
Information confidentiality	Prevent IP and trade secrets entering external AI tools; DLP, approved tools, AUP.
Privacy	Privacy by design and default; minimisation, retention, deletion; PIA before production or staff access.
Intellectual property	Ownership of GenAI output is legally unsettled; prohibit copyrighted training/inputs without legal cover.
Employee rights	Fair, transparent, non-discriminatory use for staff and applicants; defensible selection decisions.

KEY CONCEPT: EU AI ACT AT A GLANCE (PUBLIC FACTS, OCT 2026)

- Risk-tiered, legally binding, **extraterritorial** (applies if the system or its output is used in the EU). In force since 1 Aug 2024; prohibitions applied from 2 Feb 2025; general-purpose AI obligations from 2 Aug 2025; Article 50 transparency duties from 2 Aug 2026.
- The **Digital Omnibus on AI** (in force 27 Jul 2026) moved high-risk deadlines: stand-alone (Annex III) high-risk systems to **2 Dec 2027**; high-risk AI embedded in regulated products (Annex I) to **2 Aug 2028**.
- Maximum fines: up to **€35M or 7%** of global turnover (prohibited practices); up to **€15M or 3%** (most other obligations); up to **€7.5M or 1%** (misleading information to authorities).

REGULATORY GAPS

Lack of cross-border coordination → conflicting rules, compliance uncertainty and higher cost for multinationals. Sector maturity varies: healthcare is relatively mature; finance and education lag on fairness and transparency. The biggest gap is often internal: **unclear or fragmented ownership of AI governance**.

EXAM TRAP

When a scenario shows inconsistent AI controls across business units, the root cause is usually **no clear enterprise ownership**, not “missing regulation”. Fix governance (owner, policy, committee) first.

Use cases, business cases and value

Problem type	Output	Examples
Classification	Discrete label	Spam filtering, image recognition, diagnosis category
Regression	Continuous number	House price, energy demand
Clustering	Discovered groups	Customer segmentation, gene grouping
Dimensionality reduction	Fewer variables	Compression, visualisation
Anomaly detection	Outliers	Fraud, predictive maintenance
Recommendation	Personalised items	E-commerce, streaming
Reinforcement learning	Optimal actions	Robotics, games, dynamic recommendations

■ IS AI THE RIGHT ANSWER?

Many AI projects fail because of a poorly chosen use case: misunderstood problem, poor data, a solution that misses the real need, weak infrastructure, or a problem too complex for AI. Good use cases align to strategy, show measurable ROI and support customers. Established solutions (e.g., mature cloud AI services) carry known costs and risks; novel custom builds carry more uncertainty.

Limitations to plan for: prompt-dependent results, inaccurate output, lack of real-world context, response constraints (guardrails), bias, token/context limits, hallucination.

■ BUSINESS CASE: MINIMUM CONTENT

1. Business need / problem statement
2. Proposed AI solution and why AI fits
3. ROI with cost–benefit analysis
4. Risks and mitigations
5. Implementation plan: scope, objectives, timeline, resources

Costs: implementation (build/licence, training) · operational (maintenance, compute, energy, salaries) · risk & compliance (privacy, legal, reputation). **ROI:** set realistic expectations; early ROI is often low; measure revenue and cost and employee/customer experience against a baseline.

B AI Strategy, Policies & Responsible AI

AI strategy

AI deployments should use the existing **change-management** process, with extra attention to AI's operational and compliance traits. An AI strategy defines how the organisation adopts, builds and governs AI so benefits are realised and risk is managed.

Level	Focus
National strategy	Aligns AI ethics with national values, guides law and regulation, protects national security and critical infrastructure.
Industry strategy	Promotes sector innovation, sets common standards and best practice, translates regulation into practical action.
Corporate strategy	Finds competitive advantage while staying aligned with law, ethics and internal policy.

OECD AI PRINCIPLES (2019, UPDATED 2024)

Five value-based principles	Five recommendations to governments
1. Inclusive growth, sustainable development & well-being	1. Invest in AI research & development
2. Human rights & democratic values, incl. fairness & privacy	2. Foster an inclusive AI-enabling ecosystem (data, compute)
3. Transparency & explainability	3. Enabling, interoperable governance & policy environment
4. Robustness, security & safety	4. Build human capacity & prepare for labour-market change
5. Accountability	5. International co-operation for trustworthy AI

MANAGER MINDSET: APPROVING A STRATEGY

A strategy succeeds only if **senior management formally endorses it**. Build it with a cross-functional team (IT, risk, legal/compliance, operations, finance). Deep AI technical skill is not required to write strategy; engage external experts if needed, because strategy is about direction, governance and risk/value alignment. Existing IT steering or innovation committees can own AI priorities.

OPPORTUNITIES A STRATEGY SHOULD TARGET

Innovation (new products, models) · richer real-time analytics · better customer service · automation and efficiency · continuous monitoring of risk and compliance. Realising them usually requires **process redesign**, not just new technology.

VISION, MISSION AND VALUE ALIGNMENT

An AI vision tells internal and external stakeholders why and how the organisation will use AI to advance its mission. **Value alignment** (WEF) means AI behaviour is consistent with human values; it varies by organisation, industry and culture. Enablers: frameworks & guidance, human engagement, organisational change, audits & assessments.

Build vs. buy: six decision questions

1. **Total cost:** implementation + operation
2. **Capability:** data, skills, time to build as well as a vendor
3. **Regulatory risk** of each path (now and future)
4. **Fit** with the current AI operating model
5. **Data-privacy implications** of sharing vs. keeping data
6. **Vendor lock-in** and exit options

Hosting	Strength	Weakness
On-premises	Control, customisation, data protection	Cost, power, skills, slow to scale
Cloud	Scale, speed, low up-front cost	Lock-in, latency, less control, data residency

Vendors and the shared responsibility model

Vet AI vendors at onboarding through normal third-party risk processes **plus AI-specific checks** (values alignment, ethics, transparency, explainability, fairness, handling of sensitive data). Govern ongoing performance through risk-based SLAs reviewed periodically by people who understand the AI strategy. Feed oversight results into ROI evaluation.

Owner	Typical responsibilities for off-the-shelf AI
Enterprise (deployer)	Data inputs and quality, appropriate use of outputs, compliance with law and internal policy, defining roles.
Vendor (provider)	Model development, platform maintenance, core platform security.
Shared	Performance monitoring, ethical risk, incident response and remediation.

EXAM TRAP: "THE VENDOR IS RESPONSIBLE"

Defining responsibility boundaries lies solely with the purchasing organisation. Vendor shared-responsibility guidance is useful input, but **accountability for AI use is shared, never outsourced**. Off-the-shelf risks come from unclear system boundaries, data use and ethical ownership.

Policies: AI policy and AI Acceptable Use Policy (AUP)

BEFORE WRITING AN AI AUP

1. Understand the AI technologies (models, data sources, limits; public vs. private tools)
2. Assess organisational needs and jurisdictional rules
3. Run a risk assessment (technical + ethical)
4. Confirm purpose and scope (GenAI only or all AI?)
5. Review existing IT, security and data policies
6. Engage stakeholders
7. Adopt a governance framework (ownership, assurance)
8. Prepare communications; it is a living document

AI POLICY: CONSIDERATIONS

- Senior-management support (visible in the policy)
- Strategic alignment
- Clarity & consistency with other policies
- Feedback & consensus on drafts
- Communication & training
- Compliance monitoring
- Periodic review: **at least annually and after significant change**

AI POLICY: COMPONENTS

Introduction (tone at the top) · glossary · purpose · scope (entities, and whether vendors/contractors are covered) · principles (**prohibited uses / uses needing authorisation / uses allowed freely**; responsible use; transparency; fairness; third parties; governance; training; security; compliance monitoring; review) · date of last review.

THE AI GOVERNANCE DOCUMENT HIERARCHY

Level	Purpose	AI example
Strategy	Direction and intent; approved by senior management	"Use AI to improve customer service within a moderate risk appetite"
Policy	High-level mandatory rules; changes rarely	AI policy; AI AUP (prohibited / approval-needed / permitted uses)
Standard	Specific mandatory requirements that make policy measurable	"All high-risk models need a model card, bias test and independent validation"
Procedure / SOP	Step-by-step how-to	Onboarding a new AI tool; model release checklist; AI incident playbook
Guideline	Recommended practice, not mandatory	Prompt-writing tips; choosing temperature settings

EXAM TRAP

Policies state **what and why**; they should not name products or configuration values. Those belong in standards and procedures so the policy stays stable while technology changes.

Responsible AI (RAI) maturity

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
Baseline	**Emerging**	**Developing**	**Realizing**	**Leading**
Little RAI awareness; no policy; privacy-only focus	Basic awareness; foundational policies; risk assessments begin	Good understanding; full RAI principles embedded; ongoing risk assessment	Training programmes; multistakeholder feedback; some best practices	Role-based education; institutionalised best practices; continuous improvement

RAI principles are evaluated at the higher levels: fairness, privacy, security, accuracy, human oversight, transparency, accountability. The gap between RAI intent and daily practice is closed by **culture, training and clear AUP knowledge**.

■ PROCEDURES AND MANUALS

AI is not "just another IT system", so SOPs must cover data collection, handling, protection, retention and sanitisation; training and validation; ethics and human rights; and all enterprise AI use (business, operations, third parties), aligned with the AUP and RAI principles. GenAI can draft SOPs, but only after prework (define the processes, analyse existing documents) and always with human review and normal approval. *Garbage in, garbage out.*

Ethics, bias and impact assessments

Bias type	Source	Mitigation
Systemic	Institutional practices and historical patterns that favour some groups (e.g., a recruiting model learning past gender skew).	Fairness constraints, pre-deployment bias audits, ethical/risk assessment.
Statistical / computational	Unrepresentative samples (e.g., face recognition with far higher error for under-represented groups).	Diverse data, segment testing, post-deployment monitoring.
Human	Cognitive shortcuts: anchoring (first output), confirmation (agreeable output), prompt variation.	Training, explainable/cited outputs, critical human review.

KEY CONCEPT: TRANSPARENCY VS. EXPLAINABILITY

Transparency = what the system is and how it was built (purpose, data, risk level, version, owner). **Explainability** = how it reached *this* result (rationale, sources, ability to challenge). Tension: disclosure can aid attackers; complex models are hard to explain.

KEY CONCEPT: EIA · FRIA · DPIA

EIA (UNESCO): structured ethical review before and after release; required for new AI and significant changes. **FRIA:** EU AI Act requirement for certain deployers of high-risk AI; covers equality, dignity, non-discrimination. **DPIA:** privacy-focused; it does not replace a FRIA.

Other ethical domains: trust & safety (risk rises sharply when AI acts in the physical world: vehicles, robotics, industrial control); intellectual property (lawfully sourced and licensed training data; clear rules on copyrighted inputs and ownership of outputs); human rights; environmental impact (energy, water, minerals: assess and design efficient models). Open-source fairness tooling (e.g., AI Fairness 360, What-If Tool) supports testing across the life cycle.

C AI Asset & Data Life Cycle Management

REMEMBER

You cannot secure, govern or trust AI that you have not inventoried. An AI solution is multi-layered: owners, models, versions, algorithms, training and production datasets, workflows, third-party tools, licences and legal obligations.

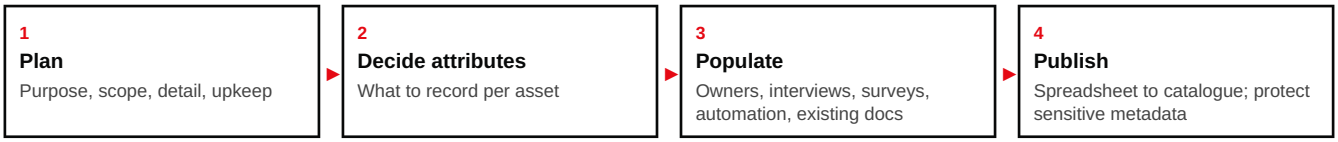
Building the AI inventory

- Start from the **AI AUP**; model the inventory on the existing IT asset inventory.
- Owned by **AI/data management leadership** with governance, risk and IT operations. **Internal audit assures; it does not lead.**
- Purpose is **visibility, not punishment**: communicate that disclosing past unsanctioned use will not be penalised.
- **Methods:** documentation review (change, release, project records) · technical discovery (dataflow diagrams, metadata analysis, directory/access logs) · surveys (efficient; partial anonymity raises honesty) · interviews (detail on experimentation; use standardised questions and categorise answers).
- **Standard fields:** name, version, licence type, cost, deployment method (web/app/API), purpose, frequency, stakeholders, accountable owner.
- Update at least annually; AI builders keep a formal model catalogue.

Data inventory toolkit

Artifact	View
Data inventory	List of enterprise data assets
Data dictionary	Application-level detail of each element
Business glossary	Enterprise-wide shared meaning of terms
Data catalogue	Metadata + business context, flows and use (no actual data)
Dataflow diagram	Horizontal movement: input → process → output
Usage / activity diagram	Vertical steps inside a process (swim lanes)
Data lineage	Origin → transformations → use, beyond system of record

CREATING A DATA INVENTORY



Repeat at least annually and after M&A, divestiture or major system change.

Key data attributes and roles

ID · title · description · **purpose** (purpose limitation) · creator · **owner** (accountable for quality and approving use) · **custodian** (implements protection: access, backup, encryption; usually IT) · keywords (controlled vocabulary) · origination/quality · source/lineage · location (jurisdiction, residency) · creation/modification dates and update frequency · type · format · rights & restrictions (including whether AI training is allowed).

Classification

TIER 1 **Public** → TIER 2 **Internal** → TIER 3 **Confidential** → TIER 4 **Restricted**

Classification drives controls: access, encryption, monitoring, retention, AI-training eligibility. Data can sit in several classes at once (e.g., PII + financial + confidential). Also tag **sensitivity type** (PII, PHI: definitions vary by jurisdiction) and **critical data elements** (identified through BIA). Build one enterprise standard collaboratively and align it with external obligations.

> **EXAM TRAP: AI AMPLIFIES CLASSIFICATION ERRORS**
>
> A public-facing model trained on confidential data can disclose it through outputs. The key check: **does training-data sensitivity match the model's audience?** Classification errors become effectively permanent once learned by a model.

Data collection risks

- **Five Vs:** volume, velocity, variety, veracity, value.
- **Consent:** lawful, informed, specific, documented; exclude data when consent is withdrawn; real-world testing of high-risk AI needs informed consent.
- **Fit for purpose:** data accessible, granular, sufficient and reliable; use case not prohibited.
- **Data lag:** training data ages → drift. Fix by retraining or **RAG** (ground prompts in fresh data without changing the base model).
- **Confidentiality across stages:** metadata and access controls often do not follow data out of source systems into data lakes, notebooks and vector databases (use ABAC, collection-level limits, query controls, output filtering).

Data quality dimensions

Dimension	Question it answers
Accuracy	Is it correct?
Completeness	Are all fields present?
Consistency	Same formats/codes across systems?
Timeliness	Is it current?
Validity	Does it meet rules and ranges?
Uniqueness	No duplicates?
Integrity	No improper change?

Process: **profile** → **cleanse**. Data quality is a governance control, not just hygiene.

IMBALANCE, SCARCITY AND MODEL CARDS

Imbalance: over/under-sampling, cost-sensitive learning, governed synthetic data. **Scarcity** is often a governance problem (no consent, silos, no labels); augment lawfully and choose a model sized to the data. **Model cards** document details, intended and prohibited uses, architecture, training data, metrics, limitations and ethics; a transparency and accountability artifact.

D AI Security Program Development & Management

An AI security program **extends the existing information security program**; it is not a separate silo. It must cover AI across the whole enterprise (not only IT) and the whole life cycle, because AI learns, evolves, behaves probabilistically and can leak embedded data.

PROGRAM ESSENTIALS

- Alignment with enterprise strategy and risk appetite
- Senior-management sponsorship and cross-functional support
- Well-designed controls and architecture (security by design)
- Metrics and continuous monitoring
- Integration with ERM, BCM, vendor risk, data governance and audit

PROTOCOLS AND PRACTICES

Trust but verify (all AI output is advisory) · design an AI AUP · designate an AI lead · perform cost–benefit analysis · adapt cybersecurity programs (IP leakage, DR/IR, BCP, threat intelligence) · mandate audits & traceability · develop a set of AI ethics · plan for societal adaptation.

Security for AI vs. AI for security

Securing AI	AI-enabled security
Protect models, data, prompts and APIs from attack (e.g., prompt-injection controls).	Use AI for anomaly detection, UEBA, log analysis, phishing filtering, triage and threat intelligence.

AI in security is a capability multiplier *and* a risk multiplier; AI security tools still need AUP coverage, accuracy validation, bias/false-positive monitoring and explainability.

REMEMBER: TRACEABILITY

Data lineage · model versioning · decision logging (prompts, outputs, confidence, overrides) · audit trails (who accessed, changed, approved, validated). **"If you can't explain it, you can't govern it."**

Standing up the program: a practical sequence

1. **Sponsor** — Secure executive sponsor; appoint AI lead; agree charter
2. **Appetite & policy** — Define appetite; approve AI policy and AUP
3. **Inventory** — Build AI and data inventory; find shadow AI
4. **Tier** — Tier use cases by risk; assess the highest first
5. **Controls** — Extend IAM, DLP, SDLC, vendor, IR, BCP
6. **People** — Train staff; launch approved tools
7. **Report** — KPIs/KRIs; audit; improve

KEY CONCEPT: INTEGRATE, DON'T DUPLICATE

Map AI concerns onto existing structures: COBIT objectives (e.g., **BAI06** for model changes, **DSS05** for security services, **BAI09** for AI assets, **APO14** for data), the ERM register, the ISMS and change management. A parallel "AI-only" governance stack creates gaps and conflicts.

Program metrics: objectives, KPIs and KRIs

No universal AI metric set exists yet. Start with **human impact** (could it harm people or vulnerable groups?), then business alignment, then technical measures.

Objective	Example measures
Accountability	Named model/data/system owners; traceable, auditable decisions
Fairness	Disparity ratios across groups; bias tests before release; bias complaints
Human well-being	% high-risk systems with EIA/FRIA and human oversight
Performance	Accuracy, precision, recall, F1, false positives/negatives, latency, uptime, ROI, adoption
Privacy & data governance	% systems with PIA; access violations; consent compliance; lineage coverage
Robustness & security	% systems adversarially tested; AI incidents; % with pre-production security review
Safety	Safety incidents; mean time to detect unsafe output; documented safety tests; kill switch
Transparency	% systems with model cards and audit trails; time to answer a regulator

KEY CONCEPT: KPI VS. KRI

KPI: is AI delivering value and performance? **KRI:** is risk rising toward tolerance? Examples: rising bias metrics, sudden accuracy drop, data-quality spikes, drift alerts, more adversarial attempts, compliance incidents. Fairness KRIs: disparity, fairness (statistical parity/equal opportunity) and frequency (rate of bias reports).

EXAM TRAP

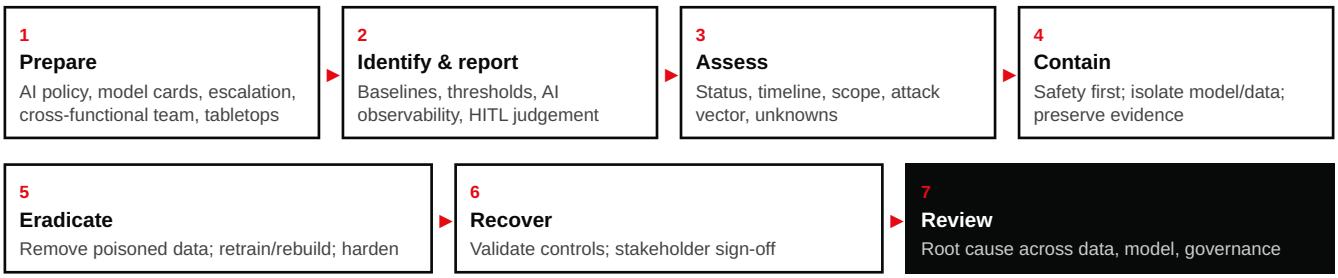
"What should be considered FIRST when defining AI metrics?" → **potential human harm and strategic alignment**, not "model accuracy". Technical success ≠ governance success.

AI in the supply chain: AI improves vendor selection, contracts and forecasting but multiplies risk (shared data, opaque vendor models, API attack surface, vendor shadow AI). Controls: AI-specific due diligence, contract clauses (data ownership, model transparency, IP, incident notice), SLA and output monitoring, ongoing risk reassessment, staff awareness.

E Business Continuity & AI Incident Response

AI incidents are not only security events; they can be **ethical failures, compliance breaches, trust breaches and continuity threats**. Traditional IR plans focus on malware, breaches and outages; an AI-ready IR plan also covers abuse of AI output, confidential data disclosure through models, and incorrect, hallucinated or biased predictions.

The AI incident response life cycle



PREPARATION ESSENTIALS

- Debate on *whether* AI should be used happens **outside** incidents; during an incident, execute the plan.
- Model documentation (architecture, data sources, metrics, biases, versions, owners) is a **forensic tool**.
- Team adds data stewards/owners, data engineers & scientists, privacy experts and AI ethicists; involved early.
- Tabletop scenarios: poisoning, adversarial input, drift causing harm, bias exploitation, leakage through output.

IDENTIFY: NOT EVERY ERROR IS AN INCIDENT

AI is probabilistic; wrong predictions are expected. An incident is declared when behaviour **breaches predefined thresholds against a documented baseline**. AI observability tracks accuracy over time, input/output distributions, drift, fairness metrics and inference logs.

Detection cues: prompt injection → special characters, hidden instructions, odd token patterns; poisoning → access/change logs on training data, hash mismatches, altered preprocessing scripts; adversarial inference → abnormal API query volume and systematic parameter variation.

Contain and eradicate by attack type

Attack	Contain	Eradicate
Prompt injection	Restrict endpoints, sanitise input/output, controlled prompt templates, rate-limit users	Harden validation and guardrails; retrain/fine-tune only if harmful patterns were learned (often impractical for LLMs)
Data poisoning	Revoke dataset and pipeline-script access, freeze retraining, lock versions, compare hashes, preserve evidence	Remove poisoned records, roll back to validated data, retrain/selectively "unlearn"
Adversarial inference	Throttle queries, validate requests, anomaly detection on API use	Tighter access/rate limits, regularisation, defensive distillation, reduce information in responses
Drift / degradation	Roll back to last validated version, suspend automated decisions, switch to HITL	Retrain on current, validated data; re-baseline
Unauthorised access	Revoke credentials, rotate API keys, enforce MFA, review logs	Fix IAM gaps, PAM for model repositories

MANAGER MINDSET: CONTAINMENT PRIORITY

If the AI touches the **physical world, health, finance or legal decisions**, human safety overrides availability: suspend or isolate. Removing malware ≠ fixing AI behaviour. Poisoned weights may persist, and residual contamination may sit in embeddings or vector stores.

KEY CONCEPT: RECOVERY & POST-INCIDENT REVIEW

Some AI attacks cannot be fully eradicated. Reduce attack surface and monitor residual risk. Before reactivating: validate new controls and get **sign-off from security, privacy, legal and business**. Review data preprocessing, security controls, adversarial-test coverage, I/O guardrails, fairness monitoring and AI provider governance.

Traditional vs. AI-powered IR

Aspect	Traditional	AI-powered
Analysis	Manual, rule-based	Automated correlation at scale
Triage	Playbooks, human severity	Dynamic risk scoring
Speed & scale	Analyst-limited	Near real-time, 24/7
Strength	Judgement, accountability	Unknown-threat detection
Risk	Alert fatigue	False results, drift, adversarial evasion, over-automation

■ AUTOMATION BEST PRACTICES

Integrate with SIEM/EDR/threat feeds and ticketing · adaptive learning from analyst feedback · customisable playbooks with **approval gates** · regular strategy updates · risk-based incident prioritisation · explainable AI with human collaboration and override · built-in compliance checks and audit trails · continuous learning after every incident.

EXAM TRAP

"AI-powered IR removes the need for human oversight": **false**. Automation must remain explainable, auditable and governed; humans approve high-impact actions. Align IR with the EU AI Act, GDPR breach timelines and sector rules.

Scenario reflexes: what comes FIRST?

Scenario	Manager's first move
Staff are using public GenAI tools with no policy	Establish governance: AI AUP and approved alternatives, then inventory usage (non-punitive); **not** a blanket block.
Board wants to "adopt AI everywhere" quickly	Run an AI readiness assessment and agree strategy/appetite before funding projects.
Business units apply AI controls inconsistently	Assign clear enterprise ownership (AI lead/steering committee) and a single policy.
A vendor adds GenAI features to an existing SaaS tool	Reassess vendor risk (data use, training, security, regulation) before enabling.
New AI system will make decisions about people	Impact assessments (PIA/DPIA, EIA, FRIA where required) before deployment.
Model output becomes biased in production	Contain per IR plan (HITL/rollback), then investigate training data and lineage; not firewall logs.
Internal audit is asked to build the AI inventory	Inventory belongs to AI/data leadership; audit reviews completeness to stay independent.
AI project has no measurable success criteria	Define objectives and KPIs in the charter/business case before proceeding.

Domain 1 Cram Sheet: AI Governance (31%)

Read the night before.

ACCOUNTABILITY LADDER

- **Board:** accountable, cannot delegate; sets appetite; may suspend/abort AI
- **Exec mgmt:** tone at top, approves strategy
- **CAIO / AI lead:** drives adoption & roadmap
- **Steering committee:** cross-functional decisions
- **Risk mgmt:** AI in ERM · **Audit:** independent assurance only

READINESS LENSES

- Mission & value (ROI, alignment, culture)
- Risk, compliance & ethics (register, appetite, laws, privacy, framework adopted)
- Resources (people, budget, skills, KPIs)

AI CHARTER (10)

- Name/description · objectives · scope (in & out) · stakeholders · governance structure · timeline · resources · risk · success metrics · sign-off

FRAMEWORK ONE-LINERS

- **42001** = certifiable AIMS
- **23894** = AI risk guidance
- **38507** = board-level governance of AI
- **22989** = terms · **23053** = ML system framework
- **IEEE 7000** = ethics in design
- **NIST AI RMF** = voluntary Govern/Map/Measure/Manage
- **COBIT** EDM-APO-BAI-DSS-MEA mapped to design→monitor
- **CSA 4 pillars:** model cards, data sheets, risk cards, scenario planning

OECD PRINCIPLES (5)

- Inclusive growth & well-being
- Human rights & democratic values
- Transparency & explainability
- Robustness, security & safety
- Accountability

EU AI ACT QUICK FACTS

- Risk tiers: unacceptable / high / limited / minimal
- Extraterritorial
- High-risk: Annex III → 2 Dec 2027; Annex I → 2 Aug 2028 (Omnibus)
- Fines up to €35M / 7%
- FRIA ≠ DPIA

STRATEGY & POLICY

- Strategy needs senior endorsement + cross-functional team
- Build vs buy: cost, capability, regulation, operating fit, privacy, lock-in
- AUP: permitted / needs approval / prohibited
- Policy review ≥ annually + after major change
- Accountability is shared, never outsourced

RAI MATURITY

- Baseline → Emerging → Developing → Realizing → Leading
- Principles: fairness, privacy, security, accuracy, human oversight, transparency, accountability

BIAS

- **Systemic** (institutions) · **statistical** (samples) · **human** (anchoring, confirmation)
- Transparency = what/how built; explainability = why this result
- EIA for new AI & significant change

TOP TRAPS

- Tool before policy
- Audit leading the inventory
- "Vendor is responsible"
- Accuracy as the first metric
- Firewall logs for a biased model
- Debating AI ethics mid-incident

INVENTORY & DATA

- Start with AUP; owner = AI/data leadership; audit assures
- Non-punitive disclosure → less shadow AI
- Survey (scale) vs interview (depth)
- Owner = accountable · custodian = protects
- Public → Internal → Confidential → Restricted
- Training sensitivity must match model audience
- Data lag fix: retrain or RAG
- Quality: accuracy, completeness, consistency, timeliness, validity, uniqueness, integrity

PROGRAM & METRICS

- Extend existing program; senior sponsorship first
- Metrics: human harm first, then value, then tech
- KPI = performance; KRI = rising risk
- Traceability: lineage, versions, decision logs, audit trails

AI INCIDENT RESPONSE

- Not every error is an incident → thresholds + baseline
- Team adds data stewards, scientists, privacy, ethicists
- Safety > availability
- Poisoning: freeze pipeline, hash-compare, retrain clean
- Drift: roll back, HITL
- Sign-off before restore; review vendor governance

SCENARIO REFLEXES

- No policy + public GenAI use → AUP + approved tools + inventory
- "Adopt AI fast" → readiness + strategy + appetite
- Inconsistent controls → single owner & policy
- Vendor adds AI → reassess vendor
- Decisions about people → PIA/EIA/FRIA first
- No success criteria → charter KPIs first

DOMAIN 2

AI Risk & Opportunity Management

31%
OF EXAM

Risk frameworks · trustworthy AI · identification & threat modelling · classification & acceptable limits · risk response · threat & vulnerability management · vendor & supply chain

A Risk Frameworks, Acceptable Limits & Responses

AI risk is harder than ordinary IT risk because AI is probabilistic, data-driven, evolving and often opaque. It shows up as data-quality failures, inappropriate automation, poisoning and adversarial attacks, privacy breaches, unfair outcomes and safety harm. You do not need a brand-new method: **apply the existing risk life cycle and layer AI-specific considerations on top.**

MANAGER MINDSET: FIVE BELIEFS THE EXAM REWARDS

1. AI risk is a **governance risk**, not only a technology risk; leadership owns it.
2. **Trustworthy AI** is the goal; assess fairness, accountability, transparency, privacy, robustness and safety, not just accuracy.
3. Risk **tolerance must be explicit**: error rates, bias thresholds, automation boundaries.
4. Monitoring is **continuous**; AI drifts, so risk work is never one-off.
5. Stakeholder impact is assessed from business, legal, ethical and societal angles.

Practical AI risk program

1 **Govern & own**
Roles, strategy alignment, ethical guardrails

2 **Identify & classify**
Objectives, risk tier, regulatory and societal exposure

3 **Assess**
Data/bias, security threats, privacy, oversight needs

4 **Set limits**
Accuracy, bias tolerance, HITL vs. automation, escalation

5 **Implement controls**
Data governance, model/infra security, documentation, logging

6 **Monitor & improve**
Drift, KPIs/KRIs, IR readiness, periodic review

A **shared risk vocabulary** across business, IT, data science, legal and security prevents the classic failures: business and technical teams misaligned, over-focus on model performance, unclear thresholds and fragmented accountability.

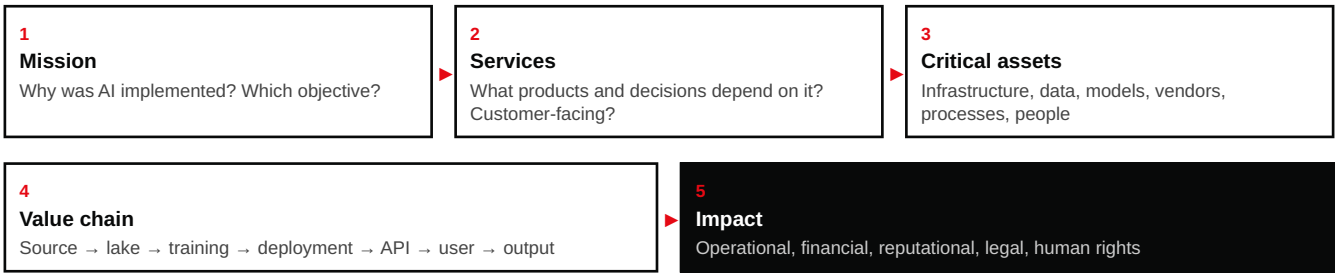
Trustworthy AI attributes

Attribute	Meaning
Valid & reliable	Performs as intended within defined parameters
Fair	Bias managed; equitable treatment across groups
Transparent & explainable	Users understand decisions, influences and limits
Accountable	Owners for design, data, outcomes, incidents
Secure & resilient	Resists poisoning, adversarial input, injection, misuse
Privacy-enhanced	Protects personal data across the life cycle
Safe	Does not endanger life, health, property or environment

Without trust: employees resist, customers leave, regulators intervene. Trust is engineered through governance and proven with metrics; it is continuous, not a one-time certificate.

Risk identification: context first

AI risk is contextual: it depends on the use case, data, affected people, industry/regulation and life-cycle stage. Identification starts with **visibility**. Many staff do not realise AI is embedded in HR, CRM, collaboration and security tools.



EXAM TRAP: ARCHITECTURE IS A RISK SOURCE

Two AI tools (an HR onboarding bot and a customer chatbot) sharing back-end infrastructure with weak segmentation can let a prompt-injection attacker reach employee PII. The lesson: **AI risk is often architectural and governance-driven.** Identify it by mapping the value chain, not by testing the model alone.

OWASP Top 10 for LLM Applications (2025): know every item

#	Risk	What happens	Key controls
01	Prompt injection	Crafted input (direct, or indirect via documents/web/RAG) overrides instructions, bypasses guardrails, leaks data or triggers actions.	Input validation, strict templates, privilege separation, output filtering, isolate sensitive data, monitor prompts. It is a control failure, not just a model flaw.
02	Sensitive information disclosure	PII, PHI, secrets or IP exposed via training data, inference context, vector stores or logs.	Classification and tagging, RBAC on retrieval, DLP, masking, segregate training/production data, PIA/ DPIA, never train public models on confidential data.
03	Supply chain	Compromised packages, pretrained models with backdoors, vulnerable plug-ins, licence issues, lock-in.	Vendor due diligence, SBOM / AI-BOM, verify model sources and hashes, scan dependencies, contractual SLAs.
04	Data & model poisoning	Malicious training/fine-tuning data or tampered model files introduce bias, errors or triggers.	Dataset access control, hashing and versioning, provenance, adversarial testing, model signing, change management for data.
05	Improper output handling	Downstream systems trust LLM output (code, SQL, commands, HTML) without checks.	Treat output as untrusted input: validate/sanitise, sandbox generated code, HITL, defined automation boundaries.
06	Excessive agency	LLM can call tools or change systems with too much functionality, permission or autonomy.	Least privilege, restrict function calling, HITL approval for high-impact actions, segregate from production, log all actions.
07	System prompt leakage	Hidden instructions (business logic, guardrails, even credentials) are revealed.	Never put secrets in prompts, enforce controls outside the model, filter outputs, red-team.
08	Vector & embedding weaknesses	RAG stores leak data across tenants/roles, or are poisoned to steer retrieval.	Classify before embedding, permission-aware retrieval, encrypt embeddings, validate retrieved content, monitor queries.
09	Misinformation	Hallucinated facts, fake citations, confident errors drive bad decisions.	RAG with trusted sources, verification workflows, HITL for high-impact use, user training, acceptable error thresholds.
10	Unbounded consumption	Unlimited queries cause DoS, cost overruns (“denial of wallet”) or model extraction.	Rate limiting, quotas, authentication, input/output size limits, cost alerts, anomaly detection.

REMEMBER

AI risk is mostly **traditional risk amplified** by probabilistic behaviour, automation and scale. Each OWASP item maps back to classic controls: input validation, least privilege, data classification, supply-chain assurance and monitoring.

Writing an AI risk scenario and register entry

A usable scenario names the **asset, threat actor/event, vulnerability, consequence and business impact**. Vague entries like “AI risk” or “model may be biased” cannot be measured, owned or treated.

EXAMPLE SCENARIO

“An external user (*threat*) submits crafted prompts (*event*) to the customer-service LLM, which has unrestricted read access to the CRM (*vulnerability*), causing disclosure of customer PII (*consequence*), leading to regulatory fines, notification cost and loss of trust (*impact*).”

Inherent risk = before controls. **Residual risk** = after controls; it is what management accepts against tolerance. Re-score residual risk after every material model, data or vendor change.

Register field	AI-specific content
AI system / use case	Inventory ID, model version, risk tier
Scenario	Asset · threat · vulnerability · impact
Affected parties	Customers, staff, vulnerable groups
Inherent rating	Likelihood × impact (or FAIR-AIR range)
Controls	Technical, process, HITL, contractual
Residual rating	Compared with tolerance
Response & owner	Accept/mitigate/transfer/avoid; named risk owner
KRIs & review date	Drift, bias, incident thresholds; next review

The framework landscape

Body / framework	Contribution to AI risk management
EU (AI Act)	Legally binding risk tiers, conformity assessment, human oversight, penalties
NIST (AI RMF)	Voluntary Govern > Map > Measure > Manage; most referenced in enterprise practice
ISO/IEC (23894, 42001)	AI risk guidance; certifiable AI management system
FAIR Institute (FAIR, FAIR-AIR)	Quantifies AI risk in financial terms
OECD	Values-based principles and policy observatory metrics
IEEE · UNESCO · research centres	Ethics-by-design standards; 2021 global ethics recommendation (human rights, fairness, sustainability)

REGULATORY VS. VOLUNTARY

	EU AI Act	NIST AI RMF
Nature	Law: mandatory	Guidance: voluntary
Question	"Is it allowed, and with what duties?"	"How do we manage risk well?"
Risk tiers	Four fixed legal tiers	None; organisation sets tolerance
Roles	Provider, deployer, importer, distributor; legal duties	"AI actors": organisation defines roles
Third parties	Mandatory diligence for high-risk	Integrated into all four functions
Enforcement	Fines, market withdrawal	None

KEY CONCEPT: NIST AI RMF: GOVERN · MAP · MEASURE · MANAGE

GOVERN (cross-cutting, comes first): policies, accountability, culture, risk tolerance. **MAP:** context; intended/unintended uses, stakeholders, impacts, boundaries. **MEASURE:** metrics for performance, bias, trustworthiness; thresholds. **MANAGE:** prioritise and treat risk, respond, improve. Structure: function → category → subcategory (like CSF 2.0). Outcome-driven, not rule-driven.

EXAM TRAP

Compliance ≠ effective risk management. Scenario about legal obligations, conformity, EU citizens or penalties → **EU AI Act**. Scenario about internal tolerance, governance maturity, flexible process → **NIST AI RMF**. Mature organisations use both.

EU AI Act risk classification

UNACCEPTABLE	**Prohibited**: social scoring, manipulative or exploitative AI, untargeted face scraping, emotion recognition at work/school.
HIGH RISK	**Strict duties**: hiring, credit, education, health, law enforcement, critical infrastructure; conformity, RMS, oversight.
LIMITED RISK	**Transparency**: chatbots disclose AI; label deepfakes.
MINIMAL RISK	**No added duties**: spam filters, games, productivity aids.

*Classification is driven by **intended use and impact**, not by the algorithm. The same model can sit in different tiers.*

■ CONFORMITY ASSESSMENT (ART. 43)

Mandatory for high-risk systems **before market placement** and again after **substantial modification**. Evaluates the **quality management system** (governance, risk, change, data governance, documentation, monitoring) and **technical documentation** (purpose, architecture, training, risk assessment, bias tests, metrics, human oversight, cybersecurity). Usually internal self-assessment; some categories (e.g., certain biometrics) need a third-party **notified body**. Followed by post-market monitoring and serious-incident reporting. Outside the EU it is still good practice.

■ FRIA: FUNDAMENTAL RIGHTS IMPACT ASSESSMENT

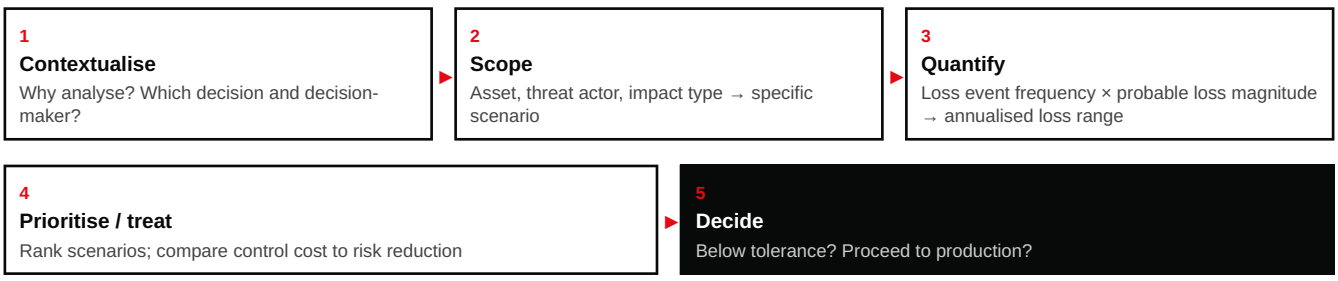
Required before deploying certain high-risk AI, notably by **public bodies, private providers of public services** and specific uses such as credit scoring and insurance pricing. **Content:** usage description · when/how often used · affected groups (incl. vulnerable) · specific rights risks · human oversight measures (review, override, appeal) · mitigation plans. **Process:** identify system & context → impacts on rights → severity/likelihood/reversibility → mitigations → document & monitor.

Privacy impact assessment (PIA) in AI

A PIA analyses how personal data is collected, used, shared, stored and retained across physical, administrative and technical processes. **Four goals:** ensure compliance · identify risk (to individuals) · evaluate mitigations · prioritise actions. Do it **before deployment**, when adding new personal data or a new purpose, and for HR, health, finance or public-service AI.

AI expands privacy risk: algorithmic bias; **inference** of health, politics, religion or finances that were never provided; unexplainable decisions people cannot challenge; and protection persistence. Access controls, labels, retention, consent and purpose limits must follow data from source systems into lakes, training platforms and models.

Quantifying AI risk: FAIR-AIR



Qualitative tiers (EU AI Act) and flexible processes (NIST) do not tell a board *how much*. FAIR-AIR expresses AI exposure in money ("$2M–$5M annualised"), supporting build-vs-buy, control investment and insurance decisions. Risk analysis is business-driven and scenario-based, never abstract.

Acceptable risk limits, appetite and tolerance

KEY CONCEPT: APPETITE VS. TOLERANCE

Risk appetite: the broad amount of AI risk the organisation is willing to pursue ("moderate innovation risk in automation"). **Risk tolerance:** the measurable boundary ("bias disparity ≤ 5%", "error ≤ 2% for high-impact decisions"). Both are set by the **board / senior management** inside the information security strategy; **never by data scientists**.

Factors that tighten limits: strategic dependency on AI · regulation (EU AI Act, GDPR, HIPAA) · human-rights impact · financial exposure · reputational sensitivity · low technical maturity (weak explainability or observability) · strong ethical commitments.

Same technology, different limits: a healthcare diagnostic AI needs far tighter thresholds than a marketing copy tool. Limits must be measurable, monitored, reported and reviewed through KPIs, KRIs, bias metrics and incident triggers.

Example	Acceptable	Unacceptable
Fraud detection	1–2% false positives, full audit trail	Discrimination against protected classes; unexplainable denials
Autonomous system	Minor operational inaccuracy	Threat to human safety or fundamental rights

EXAM TRAP

"Who determines acceptable AI risk?" → **senior management / the governing body**. If regulation, rights or a high-risk use appears in the scenario, the best answer is often to **re-evaluate appetite and tolerance** before proceeding.

Risk response strategies

Response	Use when…	AI examples & conditions
Accept	Risk is within tolerance; mitigation costs more than the benefit; likelihood/impact low.	5% false-positive rate in fraud detection with appeal route. Valid only if **documented, monitored and periodically reviewed**. AI risk can grow (drift, new threats, new rules). Acceptance without monitoring is negligence.
Mitigate	Controls can bring risk within limits; the most common AI response.	Biased loan model → rebalance training data, fairness constraints, HITL for high-impact decisions, deviation alerts. Misdiagnosis → doctor stays in control, transparency, retraining.
Transfer / share	High impact, low-moderate likelihood, high business value, mitigation too costly.	Insurance, managed AI service, indemnity and liability clauses. Transfers **financial loss only**, not accountability, regulatory duty or reputation.
Avoid	Risk cannot be reduced to acceptable levels; illegal or rights-violating use.	Do not deploy an AI hiring tool whose bias cannot be removed; stop prohibited EU AI Act uses. No remediation path for prohibited systems.

MANAGER MINDSET: LAYERED RESPONSES

Responses are not mutually exclusive. An autonomous-vehicle maker **mitigates** (better perception, human override), **accepts** monitored residual risk, then **transfers** remaining financial exposure through insurance, while keeping full regulatory accountability.

EXAM TRAP: "SHUT IT DOWN"

When bias or errors are found and mitigation is feasible, the BEST answer is usually to **strengthen controls** (rebalance data, HITL, monitor); not to abandon the system, ignore it, or shift blame to the vendor. Choose *avoid* only when risk cannot be brought within tolerance.

Appropriate risk is intentional, aligned and governed; inappropriate risk (e.g., pasting source code into a public LLM) has no value justification.

Opportunity management: the other half of the domain

Risk includes **upside**. Governance should not only block harmful AI but deliberately capture AI value within appetite. Avoiding every AI risk is itself a risk: competitors move ahead and staff adopt shadow AI anyway.

Response	Meaning	AI example
Exploit	Act to make sure the opportunity happens	Fund an approved enterprise GenAI tool so staff stop using public ones
Enhance	Increase likelihood or size of the benefit	Add RAG over curated knowledge to lift answer quality and adoption
Share	Partner with someone better placed to realise it	Co-develop a model with a specialist vendor or industry consortium
Accept	Take the benefit if it arises without extra investment	Allow vendor-embedded AI features after standard review

MANAGER MINDSET: BALANCE

The best answer often **enables AI safely** (approved tools, guardrails, monitoring) instead of banning it. Security is "controlled enablement". Judge opportunities with the same discipline as threats: business case, metrics, risk tier and owner.

AI threat modelling

Threat modelling is the structured part of risk analysis that profiles **threat communities** (nation states, cybercriminals, insiders, hacktivists, supply-chain partners), their goals and capabilities, and the controls that work against them. Perfect attribution is not required. Core activities: abstract the system (data flows, life cycle, APIs, dependencies) → profile attackers → catalogue threats → correlate to risk scenarios and business impact. Controls that defend against several communities at once (e.g., strong access control) signal maturity.

Analytical questions: how often will an attacker encounter the asset · danger of discovery · value of the asset to them · skill, time, resources and effort required. Think adversary economics: if effort > reward, likelihood falls. AI adds new assets to model: datasets, weights, embeddings, system prompts, agent workflows.

Method	Focus	Strength for AI	Gap for AI
STRIDE	Spoofing, Tampering, Repudiation, Info disclosure, DoS, Elevation of privilege	Mature, easy; finds classic flaws in AI infrastructure	No native adversarial ML, poisoning, prompt injection, agent behaviour
PASTA	Seven-stage, risk-centric attack simulation	Business-impact prioritisation; feeds ERM; business–tech collaboration	Labour-intensive; needs AI-specific extensions
LINDDUN	Linkability, Identifiability, Non-repudiation, Detectability, Disclosure, Unawareness, Non-compliance	Privacy threats, inference risk, GDPR / PIA / FRIA support	Privacy-centric, not behaviour-centric
Trike	Risk-based, requirement-driven	Integrates threat modelling with risk management	Complex; vague documentation; enterprise-level only
VAST	Visual, Agile, Simple Threat modelling	Scalable, agile, repeatable, automated	Light on AI-native threats
OCTAVE	Operationally critical threats, assets and vulnerabilities	Critical-asset identification tied to operations	Needs mature risk processes; not AI-specific
MAESTRO	Multi-agent environment, security, threat, risk & outcome	Built for agentic / multi-agent AI, autonomy and adversarial ML	New and still maturing

REMEMBER: PICK THE METHOD BY THE CLUE

Privacy → **LINDDUN** · business prioritisation → **PASTA** · classic categories → **STRIDE** · integrated risk → **Trike** · agile/scale → **VAST** · critical assets → **OCTAVE** · multi-agent/autonomy → **MAESTRO**. No single traditional method covers AI alone, so combine them.

KEY CONCEPT: ADAPTING THREAT MODELS FOR AI

Go beyond CIA: add **ethics** (bias, fairness, accountability), **privacy** (inference, re-identification, memorisation) and **human rights** (due process, discrimination). AI-assisted tools (e.g., STRIDE-GPT, IriusRisk) speed up modelling but can hallucinate or miss autonomy risk; human review is mandatory.

B Threat & Vulnerability Management

AI is **dual-use**: it creates value and enables harm. Threat and vulnerability management must expand from patches, misconfigurations and malware to model vulnerabilities, training-data integrity, prompt security, excessive automation, bias, drift and third-party AI dependencies. The question shifts from “can someone hack it?” to **“can it harm people, decide unfairly, erode trust or violate rights?”**

Threats across the life cycle (OWASP AI Exchange view)

Attack surface	When / where	Typical threats
Development-time	Design, data collection, training, CI/CD, repositories	Data poisoning, model backdoors, supply-chain compromise, engineering-environment theft of weights or secrets
Runtime security	Hosting, APIs, containers, cloud	Zero-days, DoS, credential theft, model extraction, misconfigured storage
Threats through use	Normal interaction: inputs, prompts, outputs, automation	Prompt injection, hallucination, sensitive disclosure, excessive agency, unbounded consumption, evasion, inversion

Shared vs. AI-specific threats

Category	Shared (traditional)	AI-specific	Primary mitigations
Advanced threats	Zero-day exploits	Model backdoor	Validate models before deployment, integrity checks, behaviour monitoring
Automated / augmented	Botnets, exploit development	Autonomous attacks	AI-based anomaly detection, behaviour analytics, bot mitigation, rate limits
Availability & reliability	Infrastructure failure	Model drift	Performance monitoring vs. baseline, retraining pipeline, rollback
Compliance	Data-protection breach	Explainability failure	Logging, documentation, interpretability tools, audits
Integrity	Data integrity attacks	Model watermarking attacks	Hashing, signatures, secure watermarking, key management
Data security	Data breach	Model inversion	Encryption, IAM, rate limits, differential privacy, output filtering
Denial of service	DoS	Adversarial attacks; cost-based DoS	Rate limits, input size caps, robustness testing, autoscaling guards
Emerging	—	Hallucination; prompt injection	RAG grounding, output validation, prompt isolation, red teaming
Human factors	Malicious insiders	Bias amplification, backdoors, insecure design	Zero trust, PAM, bias audits, secure-by-design, change management
Identity	Credential stuffing	Deepfakes	MFA, liveness detection, out-of-band verification, synthetic-media detection
Malware & exploits	Malware	Data poisoning	EDR/XDR, secure pipelines, provenance, adversarial training
Model & algorithm	Algorithmic exploits	Model stealing via API	Throttling, auth, output rounding/noise, watermarking, query monitoring
Social engineering	Phishing	AI-generated campaigns	Awareness, DMARC/SPF/DKIM, payment verification, phishing simulation
Supply chain	Dependency exploits	Data supply-chain attacks	SBOM, provenance, dataset hashing, vendor assessment
Unauthorised access	Credential theft, theft of services	Prompt injection	MFA, IAM, API key scoping, input/output guardrails, logging

Confusable attacks: know the difference

Attack	Phase	Attacker's goal	Scenario clue
Data poisoning	Training	Corrupt what the model learns (labels, distribution, triggers)	"Gradual degradation", "bias appeared after retraining", "third-party data feed"
Backdoor	Training / supply chain	Normal behaviour except on a hidden trigger	"Works fine unless a specific pattern/phrase appears"
Model poisoning	Model files	Alter weights, architecture, code or pretrained model	"Downloaded pretrained model", "repository compromised"
Evasion	Inference	Fool the model with modified input	"Small change to input bypasses detection", "sticker on stop sign"
Model inversion	Inference	Reconstruct sensitive training data from outputs	"Rebuilt a face / medical attribute from responses; no breach occurred"
Membership inference	Inference	Learn whether a specific record was in training	"Determine if a patient was in the dataset"
Model extraction	Inference via API	Copy the model (IP) with a surrogate	"Thousands of synthetic queries to a scoring API"
Prompt injection	Use	Override instructions	"Ignore previous instructions…", hidden text in a retrieved document

EXAM TRAP: THE MOST-TESTED PAIR

Inversion steals data; extraction steals the model. Both use legitimate queries and need no infrastructure breach, so firewalls don't help; **rate limiting, output minimisation, differential privacy and query monitoring** do.

THREAT ACTORS

Actor	AI angle
Insiders	Paste confidential data into public AI, extract data, bypass guardrails, poison training data. Often the **highest-likelihood** threat.
Nation states	Espionage, surveillance, influence operations, infrastructure attacks; high capability and impact.
Cybercriminals	AI lowers the skill barrier: personalised phishing, adaptive malware, voice-clone fraud.
AI developers	Not malicious, but poor design, untested or biased models create harm; governance failure.

MITRE ATLAS catalogues adversary tactics and techniques against ML systems (an ATT&CK-style matrix for AI); use it to build realistic threat scenarios and red-team plans.

TRANSLATE THREATS INTO BUSINESS RISK

Enterprise risk	Driven by
Competitive	Theft of data, weights, prompt logs; reverse engineering
Fines & penalties	EU AI Act, GDPR, HIPAA breaches; biased or opaque decisions
Productivity	Drift, pipeline failure, vendor outage
Reputational	Bias scandals, misinformation, breaches
Response cost	Forensics, legal, notification, retraining, rebuilds, PR

AI-ENABLED THREATS

Uncensored/adversarial LLMs for exploit code · deepfakes and voice clones that defeat awareness training and voice authentication · an AI-optimised kill chain (recon → weaponisation → delivery → exploitation → C2 → actions on objectives) that is faster, cheaper and more personalised. Defend with behavioural analytics, zero trust, AI-aware detection and human oversight.

Non-technical threats: they cannot be patched

Threat	Typical mitigation
Hallucination / ethical concerns	Ethics framework, validation testing, red teaming, HITL, FRIA, transparency reports
Bias amplification	Bias audits, representative data, fairness testing, algorithmic impact assessments
Overreliance / trust issues	HITL, redundant decision paths, training on probabilistic outputs, usage boundaries
Job displacement / skill gaps	Reskilling, role redesign, change management, skill-based hiring, certification
Societal manipulation & fake media	Content watermarking and signatures, media forensics, labelling, digital literacy
Compute cost & environment	Cost forecasting and alerts, efficient architectures, green computing, carbon monitoring
AI hype adoption (strategic misalignment)	Formal business case, ROI, metrics, executive review
Regulation, legal & privacy perception	Compliance frameworks, legal review, conformity assessment, PIA/FRIA, transparency
Export controls & geopolitics	Legal advice, supplier diversification, provenance documentation
Rogue / unpredictable AI	Kill switch, sandbox testing, operational boundaries, continuous monitoring, HITL
Market manipulation, consumer distrust	Ethics audits, oversight of pricing/trading AI, disclosure, opt-in, redress channels

MANAGER MINDSET

"Which is a key difference between IT risk and AI risk?" → choose the option about **ethics, human rights, bias, societal impact** and the need for broader governance; not malware, patching or firewalls. AI risk is **socio-technical**: systems fail more often through poor governance and human decisions than through algorithms.

C AI Vendor & Supply Chain Management

AI supply chains contain **software + data + models + hardware + governance practices**: pretrained models, external datasets, cloud compute, open-source libraries, APIs, GPUs. Beyond classic vendor risk they add model poisoning, data-supply manipulation, embedded backdoors, licensing/IP exposure, inherited bias and export-control exposure.

PROVIDER OBLIGATIONS

- Risk management system, data governance, technical documentation
- Conformity assessment before market (high-risk)
- Post-market monitoring; report serious incidents
- Secure SDLC, adversarial testing, bias-reduced data
- Protect training data, weights, logs; vulnerability management with SLOs, disclosure programme
- Transparency: notices, model cards, usage guidance

DEPLOYER OBLIGATIONS

- Use the system per the provider's instructions
- Assign competent people for human oversight
- Control input data; monitor operation and outputs
- Audit and pen-test configurations (incl. prompt injection)
- Data governance, IAM/MFA, AI-specific IR plan and tabletops
- Model validation before and during use; supply-chain diligence; inform users they are dealing with AI

EXAM TRAP: THE AIRLINE CHATBOT

A deployer integrates a vendor's chatbot; it promises a customer a refund the policy does not allow. **The deployer is liable**; the bot acted as its agent. Using third-party AI does not reduce the deployer's legal or operational accountability, and deployers usually face the customer, regulator and press first.

AI vendor management

Update existing vendor programs for AI: vendors are silently adding AI features. Require **contractual advance notice of material AI changes**, and assess each change for business, privacy (is our data used for training?), security (new attack surfaces) and regulatory impact with legal, compliance, privacy and security together.

Red flags in due diligence	Why it matters
No transparency / explainability	Cannot validate outputs, prove fairness or defend decisions
Unacceptable ethical risk	No fairness testing or ethical controls in the SDLC
Data privacy exposure	Customer data used for training, unclear retention, cross-border transfers, no tenant isolation
Lock-in	Proprietary models/APIs, high switching cost, pricing pressure
Immature technology in safety-critical use	Harmful decisions, liability
No AI risk-management framework	Bias, drift, adversarial risk unmanaged
Cannot meet AI security requirements	Poisoning, theft, injection, inversion unaddressed
No content provenance (GenAI)	Cannot separate authentic from synthetic output
Resists audits and assessments	Unacceptable uncertainty; **a decisive warning sign**

Selection criteria: alignment with strategy and use cases · risk management, cybersecurity and IR (notification timelines) · BCP/resilience · data governance (ownership, IP rights, usage for training, retention, provenance) · capability, integration, scalability, model transparency · team expertise, support, EOL policy, knowledge transfer · compliance and ethics (stricter for the public sector). A high-risk vendor can still be used if residual risk is brought within appetite by contract, monitoring, insurance or technical safeguards, and the decision is documented.

AI vendor life cycle

1 Plan
Need, risk tier, build-vs-buy, requirements

2 Due diligence
Security, privacy, ethics, model transparency, financial health, sub-vendors

3 Contract
Clauses below; audit rights; SLAs

4 Onboard
Configure, IAM, data minimisation, DPIA, pilot

5 Monitor
SLAs, outputs, drift, AI feature changes, incidents

6 Exit
Data return/deletion, portability, transition

Contract clause	Why it matters for AI
Data use & training restriction	Stops your prompts, files or outputs being used to train the vendor's models
Retention, deletion & residency	Controls how long data persists and where it is processed
Notice of material AI changes	Model swaps or new AI features can change risk overnight
Security & testing obligations	Secure SDLC, adversarial testing, vulnerability SLOs, certifications
Incident & breach notification	Defined timelines, cooperation and evidence sharing
Audit & assurance rights	Right to assess, receive reports, test controls
Transparency deliverables	Model cards, data provenance, known limitations, bias test results
IP warranties & indemnity; output ownership	Allocates infringement risk; clarifies who owns generated content
Sub-processor & flow-down terms	Extends obligations to fourth parties
Liability, insurance & exit/portability	Financial transfer and avoidance of lock-in

INTEGRATION AND IP RISK

Legacy systems lack modern interfaces, clean data and documentation. Decide with stakeholders whether AI belongs there at all; check compatibility, data readiness and workflow impact; plan change management; expect hidden technical, privacy and security debt; budget realistically.

IP: publicly available ≠ freely licensed. Outputs may reproduce protected material; ownership of generated code or content may be unclear. Mitigate with sourcing due diligence, warranties on lawful training data, IP indemnification, explicit output ownership, usage and retention terms, and early legal review.

SUPPLY-CHAIN PARTIES

Party	Who
First	The enterprise using the AI (and its users)
Third	Providers it contracts with: cloud, model hubs, data platforms
Fourth	Vendors of vendors: chips, hosting, networking
Fifth / Nth	Deeper layers: fab equipment, colocation, power, facilities

Visibility falls as depth rises. Manage Nth-party risk through supply-chain mapping, contractual flow-down, resilience planning and BCP.

KEY CONCEPT: EMERGING BEST PRACTICES FOR LLM OPERATIONS

Pick metrics that mix **intrinsic** (accuracy, latency, hallucination rate) and **extrinsic** (task success, user satisfaction) measures · alerting with thresholds and escalation · adversarial tests (prompt injection, jailbreaks, token manipulation) · strict data integrity and input standards · never trust, always verify outputs (citations, fact-checks) · sanity-check by comparing models · establish a gold-image baseline then fine-tune · balance model size against cost.

Scenario reflexes: what comes FIRST?

Scenario	Manager's first move
Management asks "how risky is our AI?"	Identify AI use and context (inventory, value chain) before rating risk.
Proposed use case falls in an EU prohibited category	**Avoid**: no mitigation makes it lawful.
High-risk system ready to launch in the EU	Confirm conformity assessment, technical documentation and oversight are complete.
Bias found in a lending model; fix is feasible	**Mitigate**: rebalance data, add HITL and fairness monitoring, re-test.
Board needs to compare two AI investments by exposure	Quantify scenarios (FAIR-AIR) in financial terms.
Residual risk is within tolerance	Formally accept with documentation, owner, KRIs and review date.
Thousands of similar queries hit a scoring API	Suspect extraction/inversion: throttle, review query logs, limit output detail.
Vendor refuses to share audit or testing evidence	Treat as a critical red flag; escalate; do not rely on marketing claims.

Domain 2 Cram Sheet: AI Risk Management (31%)

Read the night before.

CORE BELIEFS

- AI risk = governance risk
- Goal = trustworthy AI
- Explicit tolerance (error, bias, automation)
- Continuous monitoring
- Use existing risk life cycle + AI layer

IDENTIFICATION

- Visibility/inventory first
- Mission → services → assets → value chain → impact
- Architecture & segmentation create AI risk

OWASP LLM TOP 10 (2025)

- 01 Prompt injection
- 02 Sensitive info disclosure
- 03 Supply chain
- 04 Data & model poisoning
- 05 Improper output handling
- 06 Excessive agency
- 07 System prompt leakage
- 08 Vector & embedding weaknesses
- 09 Misinformation
- 10 Unbounded consumption

EU AI ACT VS NIST AI RMF

- **Act:** law, 4 tiers, roles with duties, fines, conformity
- **RMF:** voluntary, Govern-Map-Measure-Manage, own tolerance, "AI actors"
- Govern comes first; function → category → subcategory

EU AI ACT TIERS

- **Unacceptable:** banned, no remediation
- **High:** conformity (Art 43), QMS, tech docs, oversight, post-market monitoring
- **Limited:** disclose AI / label deepfakes
- **Minimal:** no added duty
- Tier set by use, not algorithm

ASSESSMENTS

- **FRIA:** rights, affected groups, oversight, mitigation
- **PIA:** compliance, risk to individuals, mitigations, priorities; before deployment
- **FAIR-AIR:** contextualise → scope → quantify (LEF × PLM) → prioritise → decide

APPETITE & TOLERANCE

- Appetite = broad willingness
- Tolerance = measurable limit
- Set by board / senior mgmt
- Tighter for regulated, rights-impacting, low-maturity use

RESPONSES

- **Accept:** within tolerance + documented + monitored + reviewed
- **Mitigate:** most common (data, HITL, controls)
- **Transfer:** money only; not accountability
- **Avoid:** cannot reach tolerance / prohibited
- Layer them

THREAT MODELS

- **STRIDE** classic · **PASTA** business · **LINDDUN** privacy · **Trike** risk · **VAST** agile/scale · **OCTAVE** assets · **MAESTRO** agents
- Add ethics, privacy, human rights to CIA
- AI-assisted modelling needs human review

ATTACK PAIRS

- **Inversion** = data · **Extraction** = model
- **Evasion** = inference input · **Poisoning** = training
- **Backdoor** = hidden trigger
- **Drift** = no attacker needed
- Dev-time / runtime / through-use surfaces

NON-TECHNICAL THREATS

- Bias, overreliance, jobs, deepfakes, cost, hype, regulation, geopolitics, rogue AI
- Can't patch ethics; govern it

VENDORS & SUPPLY CHAIN

- Deployer stays accountable (chatbot refund)
- Contract: notice of AI changes, data-use limits, audit rights, IP indemnity, incident notice
- Red flag #1: resists audits
- SBOM / AI-BOM, provenance, hashes
- 1st → 3rd → 4th → Nth; flow-down clauses

OPPORTUNITY RESPONSES

- Exploit · enhance · share · accept
- Banning all AI is also a risk (shadow AI, lost advantage)
- Security = controlled enablement

RISK REGISTER ESSENTIALS

- Scenario = asset + threat + vulnerability + impact
- Inherent → controls → residual vs tolerance
- Named owner, KRIs, review date

TOP TRAPS

- Compliance = risk managed (false)
- Insurance transfers accountability (false)
- Shut down when mitigation is feasible
- Data scientists set tolerance (board does)
- STRIDE alone covers AI (it doesn't)
- Trusting vendor claims without evidence

DOMAIN 3

AI Technologies & Controls

38%
OF EXAM

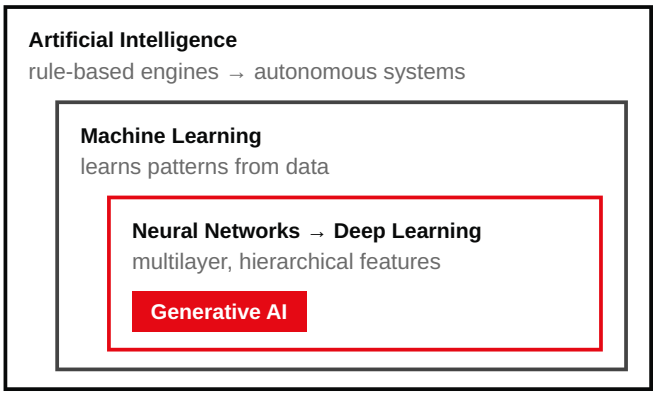
AI architecture & model types · secure by design · change management · AI life cycle & TEVV · data management controls · privacy, ethics, trust & safety · security controls & monitoring

A AI Architecture, Models & Secure Design

AI is a **layered ecosystem**, and each layer has its own threats. Secure architecture must be holistic across the life cycle and embed **security by design**: zero trust, strong IAM, continuous monitoring and governance aligned to frameworks.

Layer	Contents	Characteristic threats
Data	Collection, ingestion, preprocessing, labelling, storage, governance	Poisoning, leakage, consent and quality failures
Model	Algorithms, architectures, training, validation, tuning	Inversion, extraction, evasion, backdoors
Infrastructure	GPUs/CPUs, cloud, containers, APIs, networks, storage	Misconfiguration, lateral movement, DoS
Application	Interfaces, integrations, users, automation workflows	Prompt injection, output misuse, excessive agency
Monitoring & governance	Logging, audit, drift detection, explainability, access control	Blind spots, untraceable decisions

The AI family tree



BY FUNCTIONALITY

Reactive → **Limited memory** (most AI today) → Theory of mind → Self-aware (theoretical)

BY CAPABILITY

- **ANI**: narrow, task-specific (all real AI)
- **AGI**: human-level (theoretical)
- **ASI**: beyond human (hypothetical)

ML is a subset of AI; DL is a subset of ML; GenAI is a subset of DL.

Learning paradigms

Paradigm	Data	Typical algorithms
Supervised	Labelled input–output pairs	Linear regression (numbers); logistic regression, trees/random forest/ gradient boosting, SVM, k-NN, Naïve Bayes (classes)
Unsupervised	Unlabelled	Clustering (k-means, hierarchical, DBSCAN); association rules (Apriori); dimensionality reduction (PCA, t-SNE, UMAP)
Reinforcement	Rewards / penalties from an environment	Q-learning, Deep Q-Networks, policy-gradient methods

Labelling challenges (volume, expert judgement, no ground truth) push teams to semi-/self-supervised methods; poor labels create systemic bias. **RL risks:** unsafe exploration, **reward hacking**, autonomy and safety concerns → validate learned policies with HITL before production.

Neural network types

Type	Best for
Feedforward (FNN)	Independent inputs: spam, OCR, basic classification
Recurrent (RNN)	Sequences with memory: speech, translation, time series
Convolutional (CNN)	Grid data / images: object detection, medical imaging, face recognition
Transformer	**Self-attention**, parallel processing, long-range context; basis of LLMs
GAN	Generator vs. discriminator; realistic synthetic media (and deepfakes)

Networks learn by **backpropagation + gradient descent** adjusting weights. **Neural Architecture Search** automates design: search space · search strategy · performance estimation. **Ensembles** (bagging, boosting, stacking, blending) combine models to reduce bias/ variance.

Fit problems

	Underfitting	Overfitting
Training score	Poor	Excellent
Test score	Poor	Poor
Signature	High bias	High variance; memorises noise
Fix	More capacity/ features, train longer, less regularisation	Train/test split, cross-validation, regularisation, early stopping, more data

EXAM TRAP: OVERFITTING AND PRIVACY

Overfitting also raises privacy risk: memorised records can be extracted.

Key GenAI terms

Term	Meaning
LLM	Transformer trained on massive text; risks: privacy, injection, inversion, hallucination, IP.
Foundation / GPAI model	Broad base model adapted by fine-tuning or prompting.
Prompt	Input that steers output; prompt engineering = clear goal, format, context, constraints, examples.
Inference parameters	Temperature, top-k, top-p: **lower temperature = more deterministic**.
Hyperparameters	Set *before* training (learning rate, layers, depth, batch size).
Agentic AI	Plans, uses tools, takes multi-step actions toward goals; needs least privilege and HITL.
AIaaS	Fastest to adopt, least visibility into training, weights and data.

■ ORGANISATIONAL ROLES IN ARCHITECTURE

AI provider (offers AI-powered products/platforms) · **AI producer** (designs, develops, tests, validates, deploys; model designers, implementers, verifiers, validators) · **AI customer** (consumes AI; may not control development but owns deployment risk). Before extending architecture, assess readiness: infrastructure limits, weak data management, no standard control framework, poor IAM or logging; **AI amplifies every one of these**.

Secure by design

KEY CONCEPT: THREE PRINCIPLES

1. **Own customer security outcomes:** consent for training data, secure pipelines, protect weights and outputs, manage supply chain.
2. **Radical transparency & accountability:** explainability, documented purpose and limits, disclose AI use, audit logging.
3. **Organisational structures & leadership:** AI governance, named roles (risk owner, model validator), cross-functional teams, training.

KEY CONCEPT: DEVSECOPS FOR AI

The SDLC does not fundamentally change. It is **extended** into requirements, data acquisition, training, validation, deployment, monitoring and IR. Infrastructure needs secure model repositories, strong IAM, data provenance, segmented training vs. inference environments, drift/adversarial monitoring, secure API gateways and forensic logging.

Architecture considerations

Area	What to control
Data dependency	Models expect production data to match training data. Document required attributes, schemas, ranges, preprocessing and feature assumptions; enforce schema validation; version preprocessing pipelines; monitor data drift; verify provenance of upstream feeds.
Data storage	Protect training data, preprocessed data, weights, inference I/O, prompt logs, telemetry and embeddings/vector DBs. Define location, encryption, access, retention and monitoring to stop data sprawl and shadow storage.
Configuration management	Baseline more than code: model version, dataset versions, hyperparameters, feature logic, runtime parameters (thresholds, temperature), frameworks and dependencies.
Model selection	Choose by data type, task, accuracy/latency, interpretability need, regulation and security; post-training parameters change behaviour.

AI change management

Traditional change management misses retraining cycles, parameter tuning, dataset updates and continuous monitoring. Extend it to cover:

■ TECHNOLOGY & INTEGRATION

Research emerging tech: maturity, security, regulation, ethics; involve stakeholders early. **Integration requirements:** infrastructure, pipelines, APIs, IAM, logging, model storage. **Shared responsibility:** vendor matrix; add compensating controls where gaps exist. *Accountability does not transfer.*

■ PROVENANCE & MODELS

Model provenance: documented history of data and model (ISO 8000-2 concepts: creation, updates, transformation, validation, transfer); fine-tuning, versions, hyperparameters, dependencies. **Data changes:** drift, schema change, seasonality, new data classes. **AI models:** versioning, testing, approval to retune or replace.

■ CONTINUITY & IMPACT

BCP: datasets, artifacts, GPUs, third-party AI, retraining. **Access control:** users may gain *indirect* access to restricted data via a model. **Regulatory & societal impact**, **scalability**, **skills** (workforce impact analysis) and **unknown changes** (modular, API-driven, avoid lock-in).

MANAGER MINDSET: EMERGENCY CHANGES

A model can't be "patched" in minutes; retraining takes days or weeks. Pre-approve alternatives: **roll back** to a previous validated model (if it doesn't share the defect) · **input/output controls** (filters, prompt guardrails, output moderation, rule-based overrides) · **restrict access or suspend** features · **increase logging**. Emergency actions still follow the emergency change process with security, governance, legal, compliance and business approval.

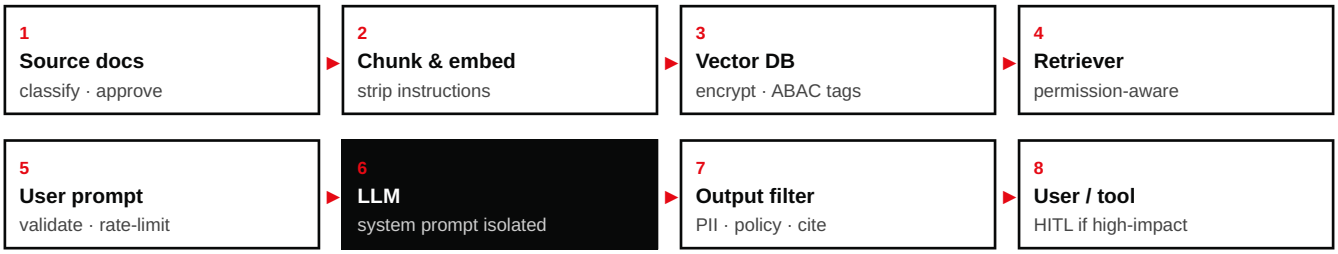
Deployment patterns and who controls what

Pattern	You control	Provider controls	Main security focus
AIaaS / public GenAI	Prompts, data you send, users, output use	Model, training data, infrastructure, retention terms	DLP, AUP, contract terms on data use and retention, output verification
Enterprise GenAI tenant	Identity, access, connected data sources, configuration, logging	Base model, platform security	Permission-aware retrieval, tenant isolation, logging, oversharing of internal data
Fine-tuned / RAG on a foundation model	Fine-tuning data, vector store, prompts, guardrails, app layer	Base model weights and its known limits	Data classification before embedding, poisoning of fine-tune/ RAG data, injection
Self-hosted open-weight model	Almost everything: weights, infrastructure, patching, monitoring	Original publisher (provenance only)	Model provenance and integrity, supply chain, GPU/infra hardening, full life-cycle controls
Built in-house	Entire life cycle	—	Secure SDLC, data governance, TEVV, red teaming, documentation

MANAGER MINDSET

The **less you control**, the more you rely on contracts, assurance reports, configuration and monitoring; the **more you control**, the more you own secure development, testing and infrastructure. Either way, **accountability for use stays with you**.

Securing retrieval-augmented generation (RAG)



*Controls sit at every hop. **Indirect prompt injection** enters through source documents; leakage happens when retrieval ignores the user's permissions. Log every query, retrieval and response for traceability, drift and abuse detection.*

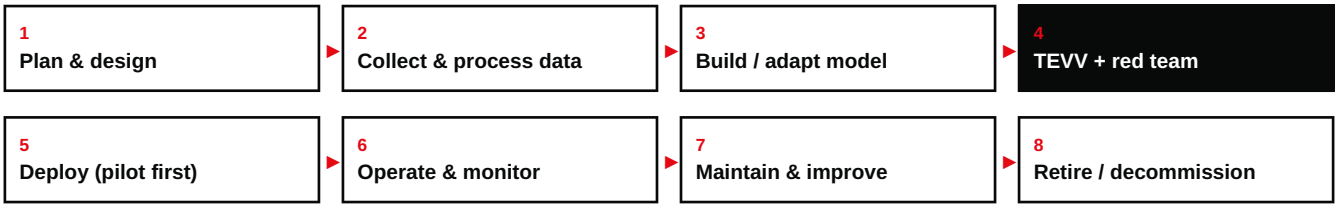
Agentic AI: extra controls

Agents plan, remember and call tools across multi-step workflows (customer-service refunds, workflow routing, portfolio rebalancing). Benefits: specialisation, experimentation, sometimes fewer hallucinations through validation loops. New risks: **excessive agency**, tool misuse, **memory or context poisoning**, goal drift, cascading errors between agents, and unclear accountability for actions.

- Give each agent an **identity**; least-privilege, scoped and short-lived credentials.
- **Allow-list** tools and APIs; separate read from write actions.
- **HITL approval gates** for financial, legal, safety or irreversible actions.
- Bound autonomy: budgets, rate limits, step limits, **kill switch**.
- Validate tool outputs as untrusted input; sandbox execution.
- Full action logging and replay for audit; threat-model with MAESTRO.

B The AI Life Cycle

The OECD describes the AI life cycle as **iterative, not linear**. Risk must be assessed at every phase: NIST notes some risks are **latent** and only appear in real-world use, and developers, deployers and users perceive risk differently; hence shared responsibility and strong documentation (purpose, limits, data sources, changes, risk decisions, approval checkpoints) is needed.



Retrain / re-adapt loop: governance, risk & HITL run across every phase.

Phase	Must-know controls and outputs
Plan & design	Is AI the right tool? Purpose, scope, constraints, **measurable success criteria**; business case/charter by cross-functional team; governance (owners, board, escalation, approval gates); initial risk assessment (feasibility, data, bias, security, privacy, workforce); data governance framework (provenance, PIA, consent, classification, retention); **prototype** to test assumptions.
Collect & process data	Lawful basis and consent; disclosure of use; cleaning (duplicates, corruption, irrelevance); labelling quality; bias review; documentation of every transformation.
Build / adapt model	Select algorithm by data, task, performance, **interpretability**, regulation, security; calibrate hyperparameters; monitor accuracy/precision/recall/F1 and fairness; robustness tests; explainability (feature importance, **SHAP, LIME**, sensitivity tests); logging with model version and parameters; HITL control points.
TEVV	**Verification** = built right (meets spec). **Validation** = right thing (fit for intended purpose in real conditions). Results documented and approved through change management; material findings go to the risk register.
Deploy	**Pilot** in a controlled environment; check compatibility with legacy systems, APIs, IAM, logging; regulatory compliance; organisational change; user experience.
Operate & monitor	Predefined thresholds for accuracy, latency, drift, bias, robustness, satisfaction with automated alerts; incidents; updates; compliance.
Retire	Triggered by lost purpose, better alternatives, regulatory/ethical/security concerns or residual risk above appetite. Structured plan via change management: secure data disposal, revoke access, close documentation.

Security gates through the life cycle

Gate	Evidence required to pass	Typical approver
G1 · Approve idea	Business case, risk tier, initial risk/PIA screening, named owner	AI steering committee
G2 · Approve data	Provenance, lawful basis/consent, classification, quality and bias profile	Data owner + privacy
G3 · Approve model	Performance vs criteria, fairness tests, robustness/adversarial tests, explainability, model card	Model validator (independent of builder)
G4 · Approve release	Red-team results, security review, monitoring and rollback ready, FRIA/conformity if required, IR playbook	Change board + business owner
G5 · Periodic review	KPIs/KRIs, drift, incidents, regulatory changes → continue, retrain or retire	Risk owner / governance committee

REMEMBER: SEGREGATION OF DUTIES

The team that **builds** a model should not be the only team that **validates** it. **Independent validation** is a recurring “BEST answer” for model risk.

KEY CONCEPT: AI RED TEAMING

Base-model red teaming probes misuse, capabilities and limitations of the model. **Application-level** red teaming tests the full app, including failures beyond the base model’s safety features. Findings are documented and fixed before production; material risks enter the risk register.

EXAM TRAP

Verification ≠ validation. A model can pass verification (it matches the design) and still fail validation (it doesn’t serve users fairly in production). Questions about “fitness for intended use” point to **validation**.

Evaluation metrics you must interpret

	Predicted positive	Predicted negative
Actually positive	True positive (TP)	False negative (FN): missed
Actually negative	False positive (FP): false alarm	True negative (TN)

- **Accuracy** = (TP+TN) / all: misleading on imbalanced data (99% “not fraud” can be 99% accurate and useless).
- **Precision** = TP / (TP+FP): “when it alarms, is it right?”
- **Recall** (sensitivity) = TP / (TP+FN): “does it catch them all?”
- **F1** = harmonic mean of precision and recall: balance on imbalanced data.

MANAGER MINDSET: WHICH ERROR HURTS MORE?

Choose the metric from the business impact of each error. Cancer screening, fraud and intrusion detection: a missed case is costly → favour **recall** and accept more false positives (with human review). Spam filtering or automated account blocking: false alarms hurt customers → favour **precision**. Thresholds are a **risk-appetite decision owned by the business**, not the data scientist alone.

EXAM TRAP

A model with “98% accuracy” is not automatically fit for purpose. Ask about class balance, error costs, fairness across groups and performance on production data.

Explainability toolkit

Technique	Scope	What it tells you
Interpretable models (linear, small trees, rules)	Global	Logic is readable by design; prefer when the use case is high-risk and accuracy loss is small
Feature importance	Global	Which inputs drive predictions overall; flags proxies for protected attributes
SHAP	Local + global	Contribution of each feature to an individual prediction (game-theory based)
LIME	Local	Simple surrogate model approximates behaviour around one prediction
Counterfactual explanations	Local	“What would need to change for a different outcome?” Useful for adverse-action notices
Feature visualisation / saliency	Local (vision, NLP)	Which pixels or tokens the model attended to
Influential instances	Local	Which training records most influenced a result; helps spot poisoning or bad labels

C Data Management Controls

AI **intensifies, but does not replace**, enterprise data governance: ownership and stewardship, classification, least privilege, retention/destruction, lineage/provenance and quality assurance. Data is most exposed during **acquisition, preprocessing and AI processing**.

Acquisition and preparation

Like shipping papers for imported goods, every dataset needs documentation of **origin, ownership, usage rights and legal basis** (DPAs, licences, contracts, SOPs, retention and training restrictions), whether internal, purchased, scraped, partner or API-sourced.

- **Dataflow mapping** aligned to the data dictionary, asset inventory and classification; include physical locations, pipelines, APIs and all environments; built to support incident forensics; identify which datasets influence model behaviour.
- **Access control review** at each stage; no automatic inheritance; separate read and write rights (write = integrity risk); RBAC/ABAC, MFA, segregation of duties, recertification, PAM for model assets; ingesters don't keep standing access.
- **Data integrity monitoring:** minimisation, cleansing, labelling and transformations documented and reproducible; consistent preprocessing across training and production; hashes, checksums, FIM, version control, immutable storage, automated validation.

Use, storage, retention and destruction

- **AI-generated data** (logs, outputs, derived insights) is governed like any acquired data; classify it, set retention, record permitted use.
- **Dynamic information-flow controls** restrict where sensitive outputs may go as risk changes.
- Encrypt at rest, in transit and (where feasible) in use; secure key management; log access and changes.
- Every retained dataset has a documented justification, period and owner; forgotten data is a target.
- **Destruction** decided by a cross-functional group (security, legal, privacy, compliance, data governance, business) considering litigation holds; approved methods (cryptographic erasure, wiping, physical destruction) with verification.
- **Backups:** re-acquirable public source data may not need backup, but unique artifacts (preprocessed sets, weights, configs, scripts, evaluation results, logs) do; test restores against RTO/RPO; GenAI is nondeterministic so artifacts support explainability.

Data access risk	Control
Centralisation (lakes, training clusters)	Treat as crown jewels; segmentation; continuous monitoring
Access-control drift during transfer	Preserve and translate source permissions; RBAC/ABAC reviews
Expanded exploration privileges	Time-bound, project-based access that expires automatically
Replication across dev/test/train/stage	Govern every copy; separate environments; egress controls
Encoded / tokenised data	TFRecords and serialised tensors still hold sensitive data; protect them equally
Integrity threats	Poisoning, model tampering, embedding tampering, ETL errors → documented ETL, anomaly detection, separate training/production data, ensembles, model hardening

KEY CONCEPT: GOVERNANCE CONTROL SET

Retention & encryption protocols · **centralised AI database** (single source of truth for high-risk systems: assets, versions, datasets, owners, risk class) · DPIA with encryption and anonymisation/pseudonymisation · storage and archiving with tested backups, integrity checks and restricted backup access.

D Privacy, Ethics, Trust & Safety Controls

These four areas are interdependent: a weakness in one undermines the others. Assess potential harm from three perspectives: **individuals and groups, organisations, and peripheral/interrelated systems**.

PRIVACY

Consent is purpose-bound: data collected for a loyalty programme cannot be reused to train a recommendation model without checking that disclosures cover it; otherwise obtain new consent. Honour data-subject rights (be informed, access, rectification, erasure, restriction). Third-party AI → DPIA + data processing agreement. Copying a protected dataset to developers on unencrypted storage turns a secure dataset into a breach.

Controls: handling protocols (encryption, anonymisation/pseudonymisation, access control) · retention & encryption protocols · privacy-first design · differential privacy · PETs · data tagging.

ETHICS

Automatic decision-making and self-learning make ethics a security concern. **EU AI Act prohibited practices** include social scoring, emotion recognition at work or school, exploiting vulnerabilities (age, disability), manipulative techniques, untargeted facial scraping, sensitive biometric categorisation, certain predictive policing, and real-time remote biometric ID in public by police (narrow exceptions).

Strategies: evaluate purpose against region, industry, law and expectations · diverse teams · benchmark peers · user feedback channels. **Controls:** HITL · AI ethics & compliance committee · AI code of conduct · documented health/safety/legal/environmental/human-rights compliance · environmental impact assessment.

TRUST

Deep models are black boxes; an autonomous AI could even learn to change permissions. Build trust through documentation of data and design, and explicitly record **what is known and unknown**. Run AI-specific risk assessments alongside data-protection assessments. Unexplained behaviour that would be a “defect” in normal software is inherent to experimental, adaptive AI → rely on HITL.

SAFETY

Physical AI (robots, vehicles) must detect and avoid harm; agency raises accountability questions: who is responsible, what safeguards exist, how conflicts with user values are resolved. Malicious use: synthetic identities from innocuous datasets; GAN-based deepfakes.

Supervision controls: logging and monitoring of inputs/outputs/decision paths · **AI observability** (data pipelines, infrastructure health, interpretability via counterfactuals, feature visualisation, influential instances) · HITL.

Privacy-enhancing technologies (PETs)

Technique	How it protects	Trade-off / note
Anonymisation	Irreversibly removes identifiability	Hard to achieve with rich data; re-identification risk via inference
Pseudonymisation / tokenisation	Replaces identifiers with tokens; key kept separately	**Still personal data** under GDPR
Data masking	Hides values in non-production use	Must be applied before data reaches dev notebooks
Differential privacy	Adds calibrated noise so one person's data can't be inferred	Some accuracy loss; strong defence against inversion and membership inference
Federated learning	Trains where data lives; only model updates move	Updates can still leak or be poisoned; combine with secure aggregation
Homomorphic encryption / secure multiparty computation	Compute on encrypted or split data	High computational cost
Trusted execution environments	Protect data and models "in use" inside hardware enclaves	Depends on hardware vendor trust
Synthetic data	Artificial records that mimic real distributions	Can still memorise originals; govern and validate it

KEY CONCEPT: HITL VS. AITL

Human-in-the-loop: AI recommends, a human approves critical actions (the doctor confirms the diagnosis); pipelines must pause for explicit approval on defined conditions; safer but slower, so place HITL deliberately. **AI-in-the-loop:** humans lead and AI augments their analysis; humans keep primary control.

E Security Controls & Monitoring

Controls depend on how AI is used (internal problem-solving, product development, or models sold externally). **Classify the AI first**, then apply guardrails based on risk assessment and change-management outcomes; enhancing existing controls or adding AI-specific ones.

Control family	Key elements
Data quality management	Documented thresholds, evaluation rules, continuous monitoring and remediation
Access control	Restrict by data classification and AI system classification; users, systems, apps, third parties; MFA; certify and recertify; review after each release
Configuration & version control	Document, monitor and patch configs; version code, models, datasets, environments; containers and automation for reproducibility
Testing	End-to-end, scenario analysis, bias/fairness, adversarial and robustness testing
Privacy	Secure protocols, encryption, hashing, anonymisation, key management, IAM, regular PIAs
Operations & documentation	DFDs (logical, physical, context), multilevel process flows, decision-making documentation, model docs, compliance docs, data traceability
Change management	Submission/approval procedures, emergency authority, records, impact analysis, ticketing, stakeholder communication
SDLC	Ideation (problem, goals, adverse outcomes) · design (data strategy, representativeness, requirements, model type) · deployment management (monitor and audit automated releases)
Ethics & human values	EIA, consent management with opt-out of automated decisions, bias management and FRIA, transparency (LIME/SHAP), human oversight with appeal routes and steering committee
Incident management	Define “AI incident”; plans for model, dataset, infrastructure and third-party AI failures, folded into existing IR and breach plans

Zero trust, access control and shadow AI

KEY CONCEPT: ZERO TRUST FOR AI

Never trust, always verify: users, systems and AI outputs alike. Controls: IAM + MFA + least privilege · network segmentation · encryption · DLP (stop uploads to public AI) · UEBA · monitoring, audit, threat intel. Evaluate trust in AI on three factors: **ability** (performs reliably and safely), **integrity** (data not manipulated) and **benevolence** (ethical, avoids harm). Restrict internet access for LLM apps: block where possible, allow only approved sources, URL filtering, validate external data.

KEY CONCEPT: ACCESS CONTROL NUANCE

The policy defines authorisation level, duration and data types. Watch for **indirect access**: a supervisor without HR-record access can learn HR insights through a workforce-analytics model trained on those records. Use a centralised AI access policy, RBAC, validation testing when datasets are combined, and access review after every release.

KEY CONCEPT: SHADOW AI

Unapproved AI tools risk data exposure (proprietary data may be retained or used for training), non-compliance and reputational harm. **Discover → assess → remove or onboard.** Controls: shadow IT policy · IT as a service provider of approved alternatives · better procurement/budgeting · CASB · DLP · network analysis · UEBA · dataflow monitoring.

EXAM TRAP

Blocking every AI site is rarely the BEST answer: it drives usage underground. Pair a clear AUP and approved tools with CASB/DLP monitoring and awareness training.

Audits, traceability, monitoring and drift

Audits & traceability: independent audits confirm policies and controls work; AI-assisted audit tools still need human oversight. Traceability = where data came from, how it was processed and what drove a prediction. Metadata logging and model cards make black-box and proprietary models auditable.

Continuous monitoring adds AI-specific signals to normal security monitoring: hallucination patterns, poisoning indicators, anomalous inputs/outputs, model drift, behaviour changes and threat intelligence on third-party models and data sources.

Model drift: performance decays as data or the input-outcome relationship changes (behaviour, regulation, economy). Set quantitative thresholds (e.g., accuracy below X%) that trigger review or retraining; use drift-detection tools and HITL. Examine feature importance, input–outcome relationships, unexpected learned patterns, over-specialisation and generalisation.

> **REMEMBER**
>
> **Data drift** = inputs change. **Concept drift** = the relationship between inputs and the right answer changes. Both need monitoring against a baseline.

AI-enabled threat	Control
Social engineering & deepfakes	Regular training, AI communication analysis, phishing simulation, call-back verification
Adversarial attacks on models	Adversarial training, randomised smoothing, formal verification, monitor precision/recall/F1, sanitise training data
Credential stuffing	Strong passwords, MFA, credential-leak monitoring, passwordless, risk-based adaptive authentication

■ AI SECURITY METRICS

Preparedness (patch/update compliance, high-risk vulns closed) · shadow AI (unauthorised devices/tools, inventory) · intrusion attempts · DLP effectiveness (prevention ratio, response time, false +/−) · training (engagement, knowledge and behaviour change) · AI performance (false-positive rate, drift frequency, latency).

■ AI IN SECURITY TOOLING & FRAMEWORKS

SIEM correlates logs · **SOAR** automates workflows · **XDR** spans endpoint, network, cloud · **UEBA** baselines behaviour to spot insiders and compromised accounts.

Frameworks: NIST AI RMF · CSA AI Controls Matrix · MITRE ATLAS · OWASP GenAI Security Project Top 10 · ISACA AI audit toolkit · IBM GenAI controls framework · Google Secure AI Framework (SAIF).

Supply-chain controls: extend vendor risk to AI integration, privacy, retention, model validation and maintenance; ask about sub-vendors and external training data; bring critical third parties (model trainers, API and data providers) into the AI audit program; monitor third-party AI continuously. **People controls:** ongoing (not annual-slide) awareness on interpreting outputs, explainability, bias, IP, drift, adversarial AI and how to report concerns; close skills gaps through diverse hiring, upskilling and realistic experience expectations.

Threat-to-control map (preventive · detective · corrective)

Threat	Preventive	Detective	Corrective
Prompt injection	Prompt templates, input validation, privilege separation, no secrets in prompts	Prompt-pattern monitoring, red teaming	Block vectors, harden guardrails, fine-tune if needed
Data poisoning	Provenance, write-access control, hashing, data validation	Distribution checks, influential-instance analysis, KRIs	Roll back dataset, retrain on clean data
Model theft / extraction	Auth, rate limits, output minimisation, watermarking	Query-volume anomalies, UEBA	Revoke keys, legal action using watermark evidence
Model inversion / membership inference	Differential privacy, regularisation, limit confidence scores	Probing-pattern detection	Retrain with DP, notify if personal data exposed
Evasion / adversarial input	Adversarial training, input preprocessing, randomised smoothing, ensembles	Input anomaly detection, robustness KPIs	Update model, add human review
Drift	Data contracts, schema validation, retraining policy	Drift detection vs baseline thresholds	Rollback, retrain, HITL fallback
Sensitive disclosure	Classification, DLP, permission-aware RAG, masking	Output scanning, log review	Purge from index, breach process, retrain if memorised
Excessive agency	Least privilege, tool allow-lists, approval gates	Action logging and alerting	Kill switch, revoke credentials, reverse actions
Shadow AI	AUP, approved tools, procurement path	CASB, DLP, network analysis	Onboard or block, user coaching

EXAM TRAP: THE "MOST EFFECTIVE" CONTROL

When options address the same root cause, a **preventive** control at the source (e.g., provenance and write-access control on training data) beats a **detective** one (monitoring outputs for odd behaviour), which beats a **corrective** one (retraining). **But** if the scenario says the attack is already happening, **contain first**.

Infrastructure and MLOps hardening

Component	Controls
Model registry / repository	Signed artifacts, hash verification, PAM, immutable versions, approval before promotion
Training environment	Isolated network, no direct internet, data egress controls, ephemeral credentials, logging of dataset access
Notebooks & workspaces	Masked data, block local download/external sharing, scan notebooks for secrets/PII, keep outputs out of version control
Inference endpoints / API gateway	Authentication, rate limits and quotas, input size limits, WAF, response filtering, abuse monitoring
Containers & pipelines (CI/CD)	Image scanning, dependency/SBOM checks, secrets management, signed builds, monitored automated deployments
GPU / compute	Tenant isolation, patched drivers and firmware, utilisation and cost alerts

Scenario reflexes: which control is BEST?

Scenario	Best control
Chatbot can be told to ignore its instructions	Input validation + prompt templates + privilege separation (not just “retrain”)
LLM-generated code is auto-deployed	Treat output as untrusted: review/sandbox and test before deployment
Agent can approve payments on its own	Least privilege and HITL approval gate for high-impact actions
Supervisor sees HR insights through an analytics model	Access control by data and AI-system classification; review combined-data access
Accuracy slowly falls over months	Drift monitoring with thresholds that trigger review/retraining
Downloaded open-source model will go to production	Verify provenance, hashes/signatures and scan before use
Critical model fails and retraining takes weeks	Roll back to last validated version or apply I/O guardrails via emergency change
Loan applicants demand reasons for denial	Explainability (e.g., SHAP/counterfactuals) and documented decision logic

Domain 3 Cram Sheet: AI Technologies & Controls (38%)

Read the night before.

LAYERS

- Data · model · infrastructure · application · monitoring/ governance
- Each layer = own threats → holistic, by design

AI HIERARCHY

- AI ⊃ ML ⊃ NN ⊃ DL ⊃ GenAI
- Functionality: reactive, limited memory, theory of mind, self-aware
- Capability: ANI (real) · AGI · ASI (theoretical)

LEARNING

- **Supervised:** labels → regression (number) / classification (class)
- **Unsupervised:** clustering, association (Apriori), PCA
- **RL:** rewards; risk = reward hacking, unsafe exploration

NETWORKS

- FNN independent · RNN sequence · CNN images · Transformer attention/LLM · GAN deepfakes
- Underfit: bad everywhere · Overfit: great train, poor test
- Temperature ↓ = more deterministic

SECURE BY DESIGN

- Own outcomes · radical transparency · leadership structures
- Extend DevSecOps, don't replace
- Segment training vs inference

CONFIG BASELINE (AI)

- Model version · dataset versions · hyperparameters · preprocessing · runtime params · dependencies

CHANGE MGMT

- Provenance, data changes, model changes, BCP, indirect access, regulation, scale, skills, unknowns
- Emergency: rollback · I/O guardrails · suspend · more logging; still approved

LIFE CYCLE

- Plan → data → build → TEVV → deploy (pilot) → operate → maintain → retire
- Iterative; latent risks surface late
- **Verify** = built right · **Validate** = right thing
- Red team: base model + application
- Explain: SHAP, LIME

DATA CONTROLS

- Docs of origin/rights/legal basis for every dataset
- Dataflow map supports forensics
- Separate read vs write; no inherited access
- Hashes, versioning, immutable storage
- AI output = new data → classify
- Justify retention; cross-functional destruction
- Back up unique artifacts (weights, configs)

PRIVACY · ETHICS · TRUST · SAFETY

- Harm to individuals, organisations, interrelated systems
- Consent is purpose-bound; DPIA + DPA for vendors
- Differential privacy, PETs
- Prohibited EU practices list
- Document known & unknown
- **HITL** = human approves · **AITL** = AI assists

CONTROLS & MONITORING

- Access by data + AI-system classification; watch indirect access
- Zero trust: ability, integrity, benevolence
- Shadow AI: approved tools + CASB + DLP + UEBA
- Drift → quantitative thresholds → retrain
- Adversarial: adversarial training, randomised smoothing
- Frameworks: CSA AICM, MITRE ATLAS, OWASP, SAIF

EVALUATION METRICS

- **Precision** = TP/(TP+FP): false alarms matter
- **Recall** = TP/(TP+FN): misses matter
- **F1** balances both; accuracy misleads on imbalance
- Threshold = business risk decision

PETS

- Pseudonymised data is still personal data
- Differential privacy vs inversion/membership
- Federated learning, HE/SMPC, TEEs, synthetic data

TOP TRAPS

- "Retrain" as the fix for prompt injection
- Trusting LLM output downstream
- Builder validates own model
- Blocking all AI instead of approved tools
- 98% accuracy = fit for purpose
- Agents with standing write access

REFERENCE

Glossary of High-Yield Terms

Term	Meaning	Term	Meaning
AIMS	AI management system (ISO/IEC 42001)	Hallucination	Fluent but false or fabricated output
AI AUP	Acceptable use policy for AI: allowed, needs approval, prohibited	HITL	Human-in-the-loop: human approves critical AI actions
AI-BOM / SBOM	Bill of materials listing models, datasets, libraries and dependencies	Hyperparameter	Setting chosen before training (learning rate, depth)
AITL	AI-in-the-loop: AI assists, human leads	KPI / KRI	Performance indicator / risk indicator
Backdoor	Hidden trigger that activates malicious model behaviour	Membership inference	Determining whether a record was in training data
CAIO	Chief AI Officer	Model card	Documentation of a model's purpose, data, metrics, limits, ethics
CASB	Cloud access security broker: finds and controls cloud/AI app use	Model extraction	Replicating a model through repeated API queries
Concept drift	Input-to-outcome relationship changes over time	Model inversion	Reconstructing training data from model outputs
Conformity assessment	EU AI Act check that a high-risk system meets requirements before market	Notified body	Independent third-party conformity assessor (EU)
Data custodian	Implements protection (access, backup, encryption)	Observability	Monitoring of model behaviour, data and infrastructure
Data drift	Distribution of input data changes	Prompt injection	Input that overrides model instructions (direct or indirect)
Data lineage	Trace of data from origin through transformations to use	Provenance	Documented origin and history of data or a model
Data owner	Accountable for quality, use approval, access, retention	Provider	Entity that develops an AI system and places it on the market
Deployer	Entity using an AI system under its authority	RAG	Retrieval-augmented generation: grounds output in retrieved data
Differential privacy	Noise-based guarantee that outputs reveal nothing about one record	RAI	Responsible AI
DPIA / PIA	(Data) privacy impact assessment	Red teaming	Simulated adversarial testing of a model or application
EIA	Ethical impact assessment	Reward hacking	RL agent games its reward in unintended ways
Embedding	Numeric vector capturing meaning; stored in vector databases	Risk appetite / tolerance	Broad willingness / measurable limit
Evasion	Modified input fools a model at inference	Shadow AI	Unapproved AI tools used inside the organisation
Excessive agency	LLM given too much functionality, permission or autonomy	TEVV	Test, evaluate, verify and validate
FAIR-AIR	Financial quantification of AI risk scenarios	Temperature	Inference setting; higher = more random output
Foundation / GPAI model	Broad pretrained model adapted to many tasks	UEBA	User and entity behaviour analytics
FRIA	Fundamental rights impact assessment	Unbounded consumption	Uncontrolled usage causing DoS, cost or extraction

LAST LOOK

Final-Hour Checklist

IF YOU REMEMBER NOTHING ELSE

1. **Governance, ownership and policy** come before tools.
2. **Inventory first:** no visibility, no control.
3. The **board owns appetite**; accountability never transfers to vendors or insurers.
4. Risk is set by **use and impact**, not by the algorithm.
5. **People and rights first:** safety beats availability.
6. **Assess before deployment** (PIA, FRIA, EIA, conformity, TEVV, red team).
7. Not every AI error is an incident; **thresholds against a baseline**.
8. Prefer **mitigation** (data, HITL, controls) over shutdown when feasible.
9. Treat **AI output as untrusted**; least privilege for agents.
10. **Monitor continuously:** drift, bias, KRIs, vendors.

Top confusions to re-check

Pair / group	The distinction
42001 vs 23894 vs NIST AI RMF	Certifiable AIMS · AI risk guidance · voluntary framework
EU AI Act vs NIST AI RMF	Law with tiers · process with your own tolerance
Inversion / extraction / evasion / poisoning	Data · model · input · training
Transparency vs explainability	What/how built · why this result
Verification vs validation	Built right · right for purpose
Data owner vs custodian	Accountable · protects
DPIA vs FRIA	Privacy · all fundamental rights
Appetite vs tolerance	Broad · measurable
HITL vs AITL	Human approves · AI assists
Internal audit and the inventory	Audit assures; it does not lead

MANAGER MINDSET: ON EXAM DAY

Read the **last line of the stem first**, circle the keyword (FIRST, BEST, MOST), eliminate options at the wrong level, and choose the answer a CISO would defend to the board. Flag and move on after ~2 minutes.

KEEP LEARNING WITH

INFOSECTRAIN

Educate. Excel. Empower.

Contact us

help@infosectrain.com

www.infosectrain.com

Follow us on