

Template 2: AI threat and risk assessment (one workbook per system)

Score each ISO/IEC 27090 threat for this system. Score = Likelihood x Impact. Yellow cells are inputs.

System ID	AI-001
System name	Recruitment screening model
Risk tier (from Template 1)	High
Assessor	[name, role]
Date	2026-10-18
Next reassessment	On retraining, vendor change, new data source, new tool, incident, or within 12 months

Ref	ISO/IEC 27090 threat	Asset hit	Objective	Likelihood (1-5)	Impact (1-5)	Score	Priority (auto)	Existing controls	Gap	Treatment	Owner	Due	Status
1	Data poisoning	Training data	Integrity	3	5	15	Treat before go-live	Vendor contract; access control on data store	No provenance checks on fine-tuning set	Add provenance validation and data snapshots	Data owner	2027-03-31	Planned
2	Evasion	Model at inference	Integrity										
3	Membership inference	Training data via model	Confidentiality	2	4	8	Treat this quarter	Internal access only; authentication	No usage logging; full confidence scores returned	Enable logging; round confidence scores	Security lead	2027-03-31	Planned
4	Model inversion	Training data via model	Confidentiality										
5	Model exfiltration	Model	Confidentiality										
6	Direct model poisoning	Model	Integrity	2	5	10	Treat this quarter	Vendor assurance questionnaire	No artefact signing; no behaviour tests	Signed artefacts; trigger tests at acceptance	CIO office	2027-06-30	Planned
7	Sensitive model output	Output	Confidentiality										
8	Prompt injection	Service / input	Integrity										
9	Output containing injection attacks	Output	Integrity (downstream systems)										
B1	Direct model theft	Model	Confidentiality										
B2	Leakage of training data	Training data	Confidentiality										
B3	Leakage of input data	Input	Confidentiality										

Summary	
Highest score	15
Threats scored	3
Treat before go-live	1
Treat this quarter	2
Accept and monitor	0
Treatments done	0

How to use this sheet

Copy this workbook once per AI system. Fill the yellow cells. Rows 1, 3 and 6 hold example values in blue italics; overwrite them.

Score and Priority calculate automatically. Rule: 15 or more = treat before go-live; 8 to 14 = treat within the quarter; below 8 = accept and monitor, and record the acceptance.

See the Threat catalogue sheet for what each threat means and the Scales sheet for the 1 to 5 definitions.

Annex B threats (B1 to B3) are conventional attacks in AI form; existing ISO/IEC 27002 controls apply.

Template by Shahzad Asghar, CISSP, CISM. Free ISO/IEC 27090 course and templates: shahzadasghar.com/learn/iso-27090/templates. Licensed CC BY 4.0: reuse and adapt with attribution.

ISO/IEC 27090:2026 threat catalogue (summary)

Summarised from the standard for use in the assessment. Verify wording against the purchased PDF.

Ref	Threat	Asset hit	Objective	AI type	Where in the standard	What it is	Consequences	How to detect	How to mitigate
1	Data poisoning	Training data	Integrity	Predictive and generative	Clause 6	Attacker inserts, alters or relabels records in training or fine-tuning data so the model learns wrong or hidden behaviour (for example a backdoor trigger).	Wrong decisions at scale, hidden backdoors, degraded accuracy, loss of trust, retraining cost. Retraining reopens the window.	Provenance and integrity checks on data sources; outlier analysis of training sets; behaviour tests with trigger inputs; accuracy drift after retraining. Automated detection remains an open research problem.	Control who and what can write to training pipelines; verify and contract data suppliers; keep data snapshots for forensics; sanitise and validate data; monitor post-deployment performance and roll back.
2	Evasion	Model at inference	Integrity	Mainly predictive	Clause 6	Inputs crafted to look normal to a person but push the model to a wrong output (altered image, shaped transaction, modified malware).	Controls that rely on the model silently fail: fraud, prohibited content or safety checks pass.	Out-of-distribution and anomaly detection on inputs; flag low-confidence or unstable outputs; compare across an ensemble; adversarial test cases in validation.	Adversarial training (with caution: can increase exposure to other attacks); input validation and preprocessing; human review for high-impact low-confidence cases; no single model for a critical decision.
3	Membership inference	Training data via model	Confidentiality	Predictive and generative	Clause 6	Attacker queries the model to determine whether a specific record was in the training set. Overfitted models answer differently for data they have seen.	Privacy breach without data theft; confirming a person is in a medical, HR or financial training set can be reportable.	Queries look like normal use. Log model usage; analyse query patterns for systematic probing; run periodic membership inference tests.	Differential privacy with a documented budget; limit or round confidence scores; regularisation and data minimisation; rate limits and authentication on inference endpoints.
4	Model inversion	Training data via model	Confidentiality	Predictive and generative	Clause 6	Attacker uses model outputs, often confidence scores, to reconstruct training data (faces, sensitive attributes).	Exposure of personal or proprietary data never directly accessible; damage comparable to a data breach.	Usage logs showing unusual systematic query sequences; per-client query volume and pattern anomalies; periodic inversion tests.	Restrict probability and confidence detail in outputs; differential privacy and data minimisation; strong authentication and per-client rate limiting; limit model access.
5	Model exfiltration	Model	Confidentiality	Predictive and generative	Clause 6	Model stealing: a functionally equivalent copy is built by querying the model many times, or parameters are obtained through weak access control.	Loss of intellectual property; the copy lets the attacker develop further attacks offline.	High-volume systematic query patterns from one client; API usage anomalies; watermark checks on suspected copies.	Authentication and per-client rate limiting; output perturbation and limited detail; watermarking; monitoring and contractual terms on API use.
6	Direct model poisoning	Model	Integrity	Predictive and generative	Clause 6	Model parameters or code altered in development or deployment, or a supplied pre-trained model arrives with a backdoor that survives fine-tuning.	Hidden malicious behaviour in production that passes normal testing; dependence on an untrusted supplier.	Integrity verification (hashes, signatures) on model artefacts; behaviour testing against trigger patterns; supplier assurance and provenance records.	Secure development environment with access and change control; signed, versioned artefacts from trusted sources; AI bill of materials; treat supplied models as supply-chain components.
7	Sensitive model output	Output	Confidentiality	Mainly generative	Clause 6	Model reveals memorised personal data or secrets, or confidential content from documents given at run time (retrieval sources).	Data breach through the front door: users read content they have no access to.	Scan outputs for personal data, secrets and classification markers; red-team extraction prompts; log outputs for review.	Minimise and classify data used in training and retrieval; same access control on retrieval sources as on documents; output filtering and data loss prevention; limit memorisation.

8	Prompt injection	Service / input	Integrity	Generative	Clause 6	Instructions hidden in input override intended behaviour. Direct (typed by user) or indirect (inside a web page, email or document). With tools, injected instructions can trigger actions.	Policy bypass, data exfiltration through the assistant, unauthorised actions through connected tools, reputational harm.	Monitor inputs and outputs for instruction-like content; canary tokens in system prompts; log prompt template versions; red-team indirect injection scenarios.	Treat all external content as untrusted data; separate instruction context from data context; least-privilege tool permissions and human approval for high-risk actions; output filtering and allow-lists.
9	Output containing injection attacks	Output	Integrity (downstream systems)	Generative	Clause 6	Model output carries a payload (script, database or operating system command) that a downstream system executes.	Compromise of applications and systems that consume AI output.	Output validation against expected formats; security testing of every integration consuming output; monitoring downstream systems for anomalous commands.	Treat model output as untrusted input in every consuming system; encode and sanitise output; parameterised queries; sandbox code execution; apply secure development rules to AI integrations.
B1	Direct model theft	Model	Confidentiality	Predictive and generative	Annex B	Model files or parameters copied from storage, repositories or deployment environments.	Loss of intellectual property; attacker gains a copy to study offline.	Access logs on model stores and repositories; integrity and inventory checks on artefacts.	Access control, encryption at rest, repository hygiene, secrets management (existing ISO/IEC 27002 controls).
B2	Leakage of training data	Training data	Confidentiality	Predictive and generative	Annex B	Unauthorised access to the data stores and pipelines that hold training and test sets.	Data breach of personal or proprietary data.	Access monitoring on data stores; data loss prevention; classification audits.	Classification, access control, separation of environments, monitoring (existing ISO/IEC 27002 controls).
B3	Leakage of input data	Input	Confidentiality	Generative	Annex B	User prompts and inputs exposed in logs, caches, vendor systems or shared sessions.	Exposure of what users typed, including personal or confidential content.	Review of log and cache contents; vendor audit of data handling; session isolation tests.	Log protection, retention limits, vendor contractual terms, tenant isolation (existing ISO/IEC 27002 controls).

Scoring scales

Adjust the wording to the organisation's risk methodology if one exists.

Score	Likelihood	Impact
1	Rare: no known technique applies, or attack needs privileged access we control tightly	Negligible: no data exposure, no decision affected, recoverable within hours
2	Unlikely: technique exists but needs effort and our exposure is internal only	Minor: limited internal data exposure or single wrong decision, recoverable within days
3	Possible: technique is documented and the system is reachable by many users	Moderate: confidential data exposure, several wrong decisions, regulatory notification possible
4	Likely: technique is common, system is external or handles valuable data	Major: personal data breach, decisions about people or money affected, reportable incident
5	Almost certain: active exploitation reported, system is public and unmonitored	Severe: large-scale breach, safety or legal consequences, sustained loss of service