

AI security risk at a glance

ISO/IEC 27090:2026 describes twelve security threats specific to AI systems. This set groups them into five business risks and one assurance programme, each with named owners and measures reported to leadership.

BRIEF	RISK	WHAT IS AT STAKE	THREATS	CONTROLS
2	Manipulation of AI training and models	Wrong decisions at scale that pass routine checks	T01, T06	AIC-01 to 04
3	Deception of AI at the point of decision	Forged documents or false readings accepted as genuine	T02	AIC-05, 06
4	Extraction of models and data	Lost intellectual property and disclosed personal data	T03, T04, T05	AIC-07 to 09
5	Direct theft and leakage of AI assets	Breaches through storage, suppliers and public AI tools	T07, T08, T09	AIC-17 to 19
6	Hijacked and leaking AI assistants	Unauthorised actions and disclosure between customers	T10, T11, T12	AIC-06, 10, 11, 12, 16
7	Ongoing assurance of AI controls	Controls that pass at launch and later stop working	A11	AIC-13 to 15

WHAT LEADERSHIP SHOULD KNOW

ISO/IEC 27090 adds AI-specific threats to the controls already run under ISO/IEC 27001. It builds on existing security.	Three of the twelve threats need no attacker. Misconfiguration, supplier error or ordinary use is enough.
Prompt injection cannot be reliably filtered out. Limiting what each AI system may do is the dependable control.	ISO/IEC 27090 is guidance and cannot be certified against. Certification remains ISO/IEC 27001 and ISO/IEC 42001.

WHO OWNS WHAT			
Data owner	AIC-01, 18	Application owner	AIC-06, 11, 12
Model owner	AIC-02, 05, 09, 15	API product owner; platform engineering	AIC-07, 08
ML platform lead	AIC-03, 17	CISO	AIC-19
System owner	AIC-04, 10	Red team lead	AIC-13
Service owner and on-call	AIC-16	Risk owner	AIC-14

DECISION FOR LEADERSHIP

Appoint one accountable executive for AI security and confirm the owner roles above, so every control has a named person answerable for it.

HOW TO USE THIS SET

Briefs 2 to 6 each cover one risk and Brief 7 covers assurance. Each points to the detailed cards used by delivery teams and auditors.

Manipulation of AI training and models

An attacker alters the data a model learns from, or the model file itself. The model still passes its routine checks but makes the decisions the attacker intends.

HOW IT COULD HAPPEN

FINANCIAL SERVICES

A fraud model retrains on disputes raised through the customer portal. Crafted disputes teach it that the attacker's transactions are legitimate.

GOVERNMENT

An open-weight model adopted for case triage runs concealed code when loaded, giving the attacker access to the development environment.

WHY IT MATTERS

Wrong decisions look normal, so losses build before anyone notices.

Without data lineage, retraining from a clean copy can take weeks.

EU AI Act Article 15(5) requires high-risk systems to resist poisoning.

WE ARE EXPOSED IF ANY OF THESE APPLY

☐ We retrain on data outsiders can influence

☐ We use third-party or open-weight models

☐ Models decide without human review

HOW WE MANAGE IT

CONTROL	WHAT IT MEANS	OWNER	REPORTED TO LEADERSHIP
AIC-01	Every dataset's source is recorded and checked for changes before training	Data owner	Live models with full data lineage (%)
AIC-02	No model is released without passing a protected test and named sign-off	Model owner	Releases that passed the gate (%)
AIC-03	Only signed and scanned model files can load	ML platform lead	Live models with a verified signature (%)
AIC-04	Each AI system has an inventory of its models, data and software	System owner	Live systems with a current inventory (%)

QUESTIONS FOR THE TEAM

If our training data had been tampered with last month, how would we know?

Can we list every external model and dataset in production today?

Could a downloaded model file run code on our systems?

A STRONG ANSWER

Datasets are fingerprint-checked and a protected test blocks bad releases.

Yes, from an inventory produced automatically at build.

No. Only signed files in safe formats can load.

DECISION FOR LEADERSHIP

Agree which AI systems are classed as high-impact. Those systems receive the strictest form of these controls first.

IF IT HAPPENS

Roll back to the last approved model and data, trace the affected records and manage it as a security incident.

Deception of AI at the point of decision

A working model is given inputs altered, often invisibly to a person, so that it reaches the wrong conclusion. Nothing in the model changes. The attack lies entirely in what is submitted.

HOW IT COULD HAPPEN

FINANCIAL SERVICES

A document classifier used for identity checks accepts a forged passport after small pixel changes push it over the threshold for a genuine document.

CRITICAL NATIONAL INFRASTRUCTURE

An anomaly detection model on an operational network receives carefully shaped traffic that keeps an intrusion just below its alert threshold.

WHY IT MATTERS

- A model can score well in normal testing and still be easy to deceive. Accuracy figures do not reveal this.
- Once the threshold is found, the same technique can be repeated cheaply and at volume.
- EU AI Act Article 15(5) requires high-risk systems to resist adversarial examples and evasion.

WE ARE EXPOSED IF ANY OF THESE APPLY

☐ A model accepts or rejects submissions automatically

☐ Submitters can see scores or reasons

☐ One party can make many attempts

HOW WE MANAGE IT

CONTROL	WHAT IT MEANS	OWNER	REPORTED TO LEADERSHIP
AIC-05	High-impact models are tested against deliberate manipulation before release and after each retrain	Model owner with the red team lead	Attack success rate at the agreed test level
AIC-06	Unusual or manipulated inputs are detected, logged and sent for review	Application owner with the SOC	Detection rate on red team cases, against false alarms

QUESTIONS FOR THE TEAM

When was our highest-impact decision model last tested against deliberate manipulation?

Do applicants or users see our confidence scores?

What happens to a decision that lands just inside the threshold?

A STRONG ANSWER

In the last twelve months, after its latest retrain, against a defined level of attack.

No. External interfaces return the decision only.

It goes to a person for review before it takes effect.

DECISION FOR LEADERSHIP

- Set the tolerable attack success rate for each high-impact model, in proportion to the harm a wrong decision would cause.
- IF IT HAPPENS
- Send flagged inputs to manual review, tighten thresholds or add a second check, and add confirmed attack samples to the test suite.

CONFIDENTIALITY

3 THREATS · 3 CONTROLS

Extraction of models and data

Using ordinary access, an attacker sends many queries and studies the answers to rebuild the model or to learn about the people in its training data.

HOW IT COULD HAPPEN

HEALTHCARE A risk model trained on one clinic's patients reveals, through its confidence scores, that a named person was a patient.	FINANCIAL SERVICES A competitor scripts millions of calls to a pricing model and trains a copy that reproduces most of its decisions.	GOVERNMENT A facial recognition service returns confidence scores that let an attacker rebuild an enrolled person's face.
---	---	---

WHY IT MATTERS

The attack uses legitimate access, so login controls alone will not stop it.	A copied model loses its value and can be studied offline to plan attacks.	Revealing who is in a dataset can be a personal data breach under UK and EU GDPR.
--	--	---

WE ARE EXPOSED IF ANY OF THESE APPLY

☐ Outsiders can query a model through an app or API

☐ Responses include scores as well as the decision

☐ A model was trained on personal or health data

HOW WE MANAGE IT

CONTROL	WHAT IT MEANS	OWNER	REPORTED TO LEADERSHIP
AIC-07	External interfaces return the minimum detail needed, with no raw scores by default	API product owner	Endpoints exposing detailed scores (target zero)
AIC-08	Every model endpoint has per-user quotas and alerts for systematic querying	Platform engineering with the SOC	Time to detect extraction in exercises
AIC-09	Models trained on personal data are tested for what they reveal, with stronger privacy protection where needed	Model owner with the DPO	Privacy leakage on sensitive models

QUESTIONS FOR THE TEAM

Which models can outsiders query, and what detail do they receive?

Would we notice one account sending a hundred times its usual volume?

Has each model trained on personal data been tested for what it reveals?

A STRONG ANSWER

A current list, each endpoint returning a decision or a band only.

Yes. A quota caps it and an alert reaches the SOC.

Yes, against a threshold agreed with the DPO, with the result recorded.

DECISION FOR LEADERSHIP Stronger privacy protection can reduce accuracy. Confirm the risk owner decides and records that trade-off for each model trained on personal data.	IF IT HAPPENS Revoke the access keys involved, preserve logs as evidence, reduce output detail at once and assess personal data exposure with the DPO.
---	--

CONFIDENTIALITY

3 THREATS · 3 CONTROLS

Direct theft and leakage of AI assets

Model files, training data and the information staff send to AI services are exposed where they are stored or processed. Two of the three threats need no attacker at all.

HOW IT COULD HAPPEN

CRITICAL INFRASTRUCTURE A forecasting model's weights sit in a public storage bucket and are copied by a scanner within hours.	HEALTHCARE A labelling contractor holding patient call transcripts, copied out for training, is breached.	GOVERNMENT Staff paste case files into a public chatbot whose provider keeps prompts and may review them.
--	---	---

WHY IT MATTERS

ISO/IEC 27090 treats securing the stack against direct theft as an efficient first investment.	A stolen model shows an attacker exactly how it works, which makes other attacks easier.	Leaked personal data triggers UK and EU GDPR duties, including 72-hour notification where required.
--	--	---

WE ARE EXPOSED IF ANY OF THESE APPLY

☐ Model files or training data sit in cloud storage or notebooks

☐ A supplier labels or prepares our data

☐ Staff can reach public AI tools at work

HOW WE MANAGE IT

CONTROL	WHAT IT MEANS	OWNER	REPORTED TO LEADERSHIP
AIC-17	Model files are classed as confidential, encrypted and readable only by named systems	ML platform lead	Model storage open to outsiders (target zero)
AIC-18	Training data copies carry their source's protection, including at suppliers, and are deleted on schedule	Data owner with procurement	Datasets held past retention (target zero)
AIC-19	Staff use approved AI services only, and sensitive fields are masked before data is sent	CISO with procurement	Attempts to reach unapproved AI services, by month

QUESTIONS FOR THE TEAM

Where are our model files stored, and who can read them?

Which suppliers hold copies of our training data?

What stops staff pasting customer or patient records into a public chatbot?

A STRONG ANSWER

In named, encrypted locations that only approved systems can read, with access logged.

A named list, each under an agreement with security and deletion terms.

Approved tools on enterprise terms, blocking of others and masking of sensitive fields.

DECISION FOR LEADERSHIP Approve the AI services staff may use, and the contract terms a new service must meet before it is added.	IF IT HAPPENS Revoke credentials, contain storage, preserve evidence, follow the breach procedure and request deletion from any provider involved.
---	--

Hijacked and leaking AI assistants

Assistants and agents read untrusted content and can act on it. Hidden instructions can redirect them, and loosely scoped retrieval can show one customer another's data.

HOW IT COULD HAPPEN

FINANCIAL SERVICES

A support assistant quotes another customer's ticket because retrieval is not scoped.

GOVERNMENT

An email assistant obeys hidden instructions and forwards mail to an outside address.

HEALTHCARE

A chatbot reply holds an image link that sends the conversation to an attacker.

WHY IT MATTERS

The damage matches whatever the assistant is allowed to do: send, pay, delete or amend.

Prompt injection cannot be reliably filtered out. Limiting permissions is the dependable control.

Disclosure between customers is a data breach and a direct loss of client trust.

WE ARE EXPOSED IF ANY OF THESE APPLY

- ☐ An assistant reads emails, documents or web pages
- ☐ An agent can send, pay or change records
- ☐ One assistant serves several customers

HOW WE MANAGE IT

CONTROL	WHAT IT MEANS	OWNER	REPORTED TO LEADERSHIP
AIC-10	Each agent has its own identity and minimum rights; irreversible actions need human approval	System owner with IAM	Agents with approval gates (%)
AIC-12	Retrieval applies the requester's own permissions at the data layer	Application owner	Cross-customer test failures (target zero)
AIC-11	Every system that shows or uses assistant output treats it as untrusted	Application owner	Output consumers with documented handling (%)
AIC-06	Injection attempts in inputs and retrieved content are detected and logged	Application owner with the SOC	Detection rate on red team cases
AIC-16	Every agent has a tested way to switch off its tools, retrieval or model	Service owner with on-call	Time to disable in the last test

QUESTIONS FOR THE TEAM	A STRONG ANSWER
What could each assistant do if hijacked?	Listed per agent, with irreversible steps approved.
Could one customer see another's data?	No. Retrieval checks permissions, and tests prove it.
How fast can we switch an assistant off?	Within minutes, by on-call staff, tested twice a year.

DECISION FOR LEADERSHIP

Set the actions no AI agent may take without a person approving, such as payments, external messages and deleting records.

IF IT HAPPENS

Use the kill switch, review the action logs, reverse what can be reversed and assess any data disclosed.

Ongoing assurance of AI controls

The controls in Briefs 2 to 6 only protect us while they keep working. This brief covers how we test them, record the trade-offs between them and re-check them as models change.

HOW IT COULD GO WRONG

HEALTHCARE

Privacy protection added to a diagnostic model lowers its accuracy for one patient group. Nobody recorded who accepted that trade-off.

FINANCIAL SERVICES

A provider updates the model behind a customer assistant. Protections tested at launch no longer hold, and no one re-tests them.

WHY IT MATTERS

Models drift and defences age. A control that passed at launch can quietly stop working.

ISO/IEC 27090 clause 7.2 warns that one mitigation can weaken another.

Infrastructure penetration tests rarely test the model itself.

WE ARE EXPOSED IF ANY OF THESE APPLY

- ☐ We rely on third-party models that change version
- ☐ No AI red team exercise in the last year
- ☐ Trade-offs between controls are not written down

HOW WE MANAGE IT

CONTROL	WHAT IT MEANS	OWNER	REPORTED TO LEADERSHIP
AIC-13	AI systems are red teamed before launch, on significant change and at least yearly if external	Red team lead	Findings closed within agreed timescales (%)
AIC-14	Conflicts between mitigations, such as privacy against accuracy, are assessed and approved by the risk owner	Risk owner	High-impact systems with recorded decisions (%)
AIC-15	Models and their defences are re-checked on a schedule and when drift or new attacks appear	Model owner	Models re-validated on schedule (%)

QUESTIONS FOR THE TEAM	A STRONG ANSWER
When was each high-impact system last red teamed?	Within twelve months, every finding closed or formally accepted.
Who approved our privacy and accuracy trade-offs?	The named risk owner, recorded with a review date.
What happens when a provider changes its model?	Red team cases are re-run after every major update.

DECISION FOR LEADERSHIP

Approve an annual AI assurance plan covering red team scope, the re-validation schedule and the reporting leadership will receive.

IF A CONTROL FAILS

Record the gap as a finding with an owner and a date, and keep it in leadership reporting until it is closed.