SRIDA IT PROFESSIONAL REFERENCE PUBLICATION

ISO/IEC 27090:2026

Cybersecurity for Artificial Intelligence Systems

AI Security Threats, Safeguards & Assessment Testing Procedures

DOCUMENT ID	VERSION	STATUS
SRIDA-AIS-27090-RTP-001	1.0	Final — Approved for Publication

Rajendra Bodda
Chief Information Security Officer (CISO)
Srida IT · 2026

Independent professional guidance.
Not an official ISO or IEC publication.

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT



ISO/IEC 27090:2026 — Cybersecurity for Artificial Intelligence Systems

AI Security Threats, Safeguards & Assessment Testing Procedures

SAMPLE

Prepared by

Rajendra Bodda

Chief Information Security Officer (CISO)

Published by Srida IT
SRIDA-AIS-27090-RTP-001 · Version 1.0 · October 2026

This publication is an independently developed professional guidance document prepared by Srida IT. It is not an official ISO or IEC publication, does not replace the licensed ISO/IEC 27090:2026 standard, and does not confer certification or accreditation. ISO/IEC 27090 is a guidance standard and is not itself certifiable.

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

Document Information and Control

Field	Value
Document title	ISO/IEC 27090:2026 — Cybersecurity for Artificial Intelligence Systems
Subtitle	AI Security Threats, Safeguards & Assessment Testing Procedures
Document identifier	SRIDA-AIS-27090-RTP-001
Version	1.0
Status	Final — Approved for Publication
Classification	Srida IT Professional Reference Publication
Author	Rajendra Bodda, Chief Information Security Officer (CISO)
Publisher	Srida IT
Publication year	2026
Issue date	October 2026
Reference standard	ISO/IEC 27090:2026 (guidance standard, first edition)
Related standards	ISO/IEC 27001:2022; ISO/IEC 27002:2022; ISO/IEC 42001:2023
Companion publications	SRIDA-ISMS-27001-RTP-001 (ISO/IEC 27001:2022); SRIDA-AIMS-42001-RTP-001 (ISO/IEC 42001:2023)
Next review	Annually, or on revision of the referenced standards

Version history

Version	Date	Author	Description of change
1.0	October 2026	Rajendra Bodda	First published edition, approved for publication.

Approval

Role	Name	Date
Approved for publication	Rajendra Bodda, Chief Information Security Officer (CISO)	25 October 2026

Distribution

Released for publication by Srida IT. Distribution and use are subject to the permitted-use terms in the Copyright, Legal and Usage Notices.

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

Table of Contents

Part I Foundations 14

Chapter 1 Introduction to ISO/IEC 27090:2026 15
Chapter 2 The Standard's Structure and Role 20
Chapter 3 AI Security Principles and the AI Attack Surface 25
Chapter 4 ISO/IEC 27090 and Related Standards, Frameworks and Regulation 30
Chapter 5 Terminology, Definitions and Abbreviations 35
Chapter 6 How to Use This Manual 40
Chapter 7 Assessment Methodology, Evidence and Sampling for AI Security 44

Part II Applying Information Security to AI 50

Chapter 8 AI Security Governance and Risk 51
G-01 AI security governance and accountability 52
G-02 AI system and AI asset inventory 57
G-03 AI security risk assessment 62
G-04 Threat modelling for AI systems 67
Chapter 9 Securing the AI Life Cycle 72
G-05 AI supply chain security 73
G-06 Secure AI development environments and ML pipelines 78
G-07 Data security across the AI life cycle 83
G-08 AI security testing and red teaming 88
Chapter 10 Operating, Monitoring and Responding 93
G-09 Logging, monitoring and malicious input detection 94
G-10 Throttling and abuse prevention for AI services 99
G-11 Deterioration of models and mitigations over time 104
G-12 AI security incident response and recovery 109

Part III Types of Compromise to AI Systems 114

Chapter 11 Training-Time and Model Integrity Attacks 115
6.2 Data poisoning 116
6.7 Direct model poisoning 123
Chapter 12 Inference-Time Confidentiality and Evasion Attacks 129
6.3 Evasion 130
6.4 Membership inference 136
6.5 Model exfiltration 142
6.6 Model inversion 148
6.8 Sensitive model output 154
Chapter 13 Language-Model and Application-Layer Attacks 160
6.9 Prompt injection 161
6.10 Output contains injection attacks 168

Part IV Mitigations: Srida IT AI Security Control Set 174

Chapter 14 Data and Model Protection 175
AIS-01 Training data provenance and integrity 176
AIS-02 Data validation and poisoning detection 182

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

AIS-03 Privacy-preserving training and data minimisation 187
AIS-04 Model artefact integrity and registry control 192
AIS-05 Third-party model vetting and safe model formats 197
AIS-06 Model access control and confidentiality 202
AIS-07 Adversarial robustness engineering 207
AIS-08 Output information minimisation 212

Chapter 15 Inference and Application Security 217
AIS-09 AI API authentication, rate limiting and throttling 218
AIS-10 Query pattern monitoring and extraction detection 224
AIS-11 Input validation and malicious input detection 229
AIS-12 Prompt injection defences 234
AIS-13 Output validation and safe downstream handling 239
AIS-14 Sensitive output filtering and leakage prevention 244
AIS-15 Retrieval authorisation for retrieval-augmented generation 249
AIS-16 AI agent and tool-use privilege control 254

Chapter 16 Life-Cycle and Operational Security 260
AIS-17 Secure ML pipeline and environment separation 261
AIS-18 AI supply chain assurance and AI bill of materials 266
AIS-19 AI security testing and red teaming 271
AIS-20 AI security logging and telemetry 276
AIS-21 Model and mitigation deterioration monitoring 281
AIS-22 AI security incident response and model rollback 285
AIS-23 Federated learning security 290
AIS-24 Secure retirement of AI models and data 295

Chapter 17 Implementation Roadmap and Integration with ISO/IEC 27001 and ISO/IEC 42001 300

Part V AI Security Assessment and Assurance 307
Chapter 18 Planning AI Security Assessments 308
Chapter 19 AI Red Teaming and Adversarial Testing Engagements 314
Chapter 20 Findings, Risk Rating and Remediation Verification 320
Chapter 21 Using ISO/IEC 27090 in ISO/IEC 27001 and ISO/IEC 42001 Certification Audits 325

Part VI Appendices and Templates 330
Appendix A Threat-to-Control Matrix 331
Appendix B AI Security Control Set Summary 332
Appendix C AI System Security Inventory Template 335
Appendix D AI Threat Model Template 339
Appendix E AI Security Risk Register Template 343
Appendix F AI Red-Team Rules of Engagement and Test Plan Template 346
Appendix G Assessment Workpaper Template and Completed Sample 350
Appendix H Finding Report and Remediation Tracker Template 354
Appendix I AI Security Incident Runbook Template 357
Appendix J AI Supplier Security Questionnaire 361
Appendix K AI Security Readiness Checklist 365
Appendix L References and Bibliography 369

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

CHAPTER 1

Introduction to ISO/IEC 27090:2026

Artificial intelligence systems are now part of ordinary business operations. Credit decisions, fraud screening, medical image triage, customer-support assistants, code generation and document summarisation increasingly depend on models that were trained on large data sets, are packaged as artefacts, and are exposed to users through application programming interfaces. These systems inherit every conventional information security risk of the software and infrastructure on which they run. They also introduce a class of risks that conventional controls were not designed to address: the behaviour of a model can be altered by the data it learns from, the confidentiality of training data can be eroded through the outputs a model produces, and a language model can be steered by instructions hidden in the very content it is asked to process.

ISO/IEC 27090:2026 is the international standard that addresses this class of risk. This chapter introduces the standard, explains what is and is not known about it at the time of writing, sets out the purpose and audience of this manual, and describes how the manual relates to Srida IT's earlier manuals on ISO/IEC 27001 and ISO/IEC 42001.

1.1 Why AI security needs its own guidance

An information security management system (ISMS) built on ISO/IEC 27001 protects information assets through risk assessment, risk treatment and a set of controls such as access control, secure development, logging, supplier management and incident response. Those controls remain necessary for AI systems, and most AI security failures observed in practice still begin with a conventional weakness: an exposed storage bucket holding training data, an over-privileged service account in a machine learning (ML) pipeline, an unauthenticated inference endpoint, or a third-party component pulled without verification.

AI systems nevertheless differ from conventional software in ways that matter for security:

- **Behaviour is learned, not written.** A model's decision logic is derived from data. Whoever can influence the training data, the fine-tuning data or the feedback loop can influence the behaviour of the system, often without changing a single line of code.
- **Data persists inside the model.** Model parameters can encode information about individual training records. Outputs and confidence scores can therefore disclose information that the organisation believed was protected by access controls on the original data set.
- **The model is itself a valuable asset.** Trained weights represent significant investment in data, compute and expertise, and can be approximated or copied through repeated querying as well as stolen directly.
- **Inputs are effectively code.** For generative AI and large language models (LLMs), natural-language input can change what the system does. The traditional boundary between data and instructions becomes blurred, which is the root of prompt injection.
- **Outputs flow into other systems.** Generated text, code, queries and tool calls are consumed by browsers, databases, interpreters and business workflows. Unvalidated model output can carry injection attacks into those downstream systems.
- **Behaviour is probabilistic and changes over time.** Models drift as data distributions change, and the effectiveness of mitigations such as filters and classifiers can deteriorate as attackers adapt. A control that passed testing at deployment may not be effective six months later.

These characteristics mean that an organisation can operate a well-run ISMS and still be exposed to AI-specific attacks that its risk assessment never identified and its controls never tested. ISO/IEC 27090 was developed to close that gap.

1.2 What ISO/IEC 27090:2026 is

ISO/IEC 27090:2026 was published as a first edition on 14 October 2026. It runs to 56 pages and sits within the ISO/IEC 27000 family of information security standards, developed under the joint technical committee of ISO and IEC responsible for information security, cybersecurity and privacy protection.

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

ISO's own description of the scope is that the standard provides guidance on understanding, detecting and mitigating security threats that are specific to AI systems throughout their life cycle. It is intended to complement, not replace, conventional information security as described in ISO/IEC 27001 and ISO/IEC 27002, and it states plainly that following its guidance does not guarantee immunity from attack.

Three features of the standard shape everything that follows in this manual:

1. **It is guidance.** ISO/IEC 27090 expresses recommendations using "should". It does not contain requirements, it is not a management system standard, and an organisation cannot be certified against it. Chapter 2 explains how it is used alongside the certifiable ISO/IEC 27001 and ISO/IEC 42001.
2. **It is threat-centred.** Its core, Clause 6, describes nine types of potential compromise to AI systems, from data poisoning to output that contains injection attacks. Each is treated in a common pattern of description, the failures that result from the attack, detection and mitigation.
3. **It spans the life cycle.** The standard addresses threats from inception and data collection through deployment and operation to retirement, and one of its informative annexes maps attacks to life-cycle stages and assets.

Status of the standard text in this manual

At the time this manual was prepared, the full licensed text of ISO/IEC 27090:2026 had not been available to the author. The facts about the standard stated in this manual are limited to those confirmed from multiple secondary sources: its publication date, edition and length; its scope; its guidance (non-certifiable) nature; the subject of Clauses 5, 6 and 7; the titles of the nine compromise types in 6.2 to 6.10; and the topics of the three informative annexes. Any other statement about the content of the standard reflects Srida IT's understanding and should be reconciled against the licensed standard. Readers should obtain the standard from ISO or their national standards body and reconcile this manual against it before relying on it for audit or certification purposes.

1.3 The nine types of compromise

Clause 6 of ISO/IEC 27090, "Types of potential compromise to AI systems", opens with a general sub-clause (6.1) and then describes nine compromise types. They are listed below with a short Srida IT explanation; Part III of this manual treats each in depth.

Table 1-1. The nine compromise types described in ISO/IEC 27090 Clause 6 (explanations are Srida IT's)

Sub-clause	Compromise type	Srida IT summary	Manual chapter
6.2	Data poisoning	Manipulating training, fine-tuning or feedback data so that the resulting model behaves as the attacker intends, including hidden backdoor triggers.	11
6.3	Evasion	Crafting inputs at inference time that cause a model to make an incorrect decision while appearing normal to a human observer.	12
6.4	Membership inference	Determining whether a particular record was part of a model's training data, which can itself disclose sensitive facts about a person.	12
6.5	Model exfiltration	Stealing a model, either by copying its artefacts or by approximating its behaviour through systematic querying.	12
6.6	Model inversion	Reconstructing representative features or records of the training data from a model's outputs.	12
6.7	Direct model poisoning	Tampering with model parameters, architecture or artefacts directly, rather than through the training data.	11
6.8	Sensitive model output	A model disclosing confidential, personal or otherwise restricted information in its outputs.	12
6.9	Prompt injection	Inputs, supplied directly by a user or embedded in content the model processes, that override or subvert the system's intended instructions.	13
6.10	Output contains injection attacks	Model output that carries a conventional injection payload (for example script, query or command content) into a downstream system that trusts it.	13

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

ISO/IEC 27090 COMPROMISE TYPE 6.4

Membership inference

ISO/IEC 27090 Clause 6 · 6.4 WP-T6.4

A Reference

ISO/IEC 27090:2026, Clause 6.4 — Membership inference

Membership inference is an attack in which an adversary determines whether a particular record, or data about a particular individual, was part of a model's training data. ISO/IEC 27090 Clause 6 identifies membership inference (6.4) as a type of potential compromise to AI systems; its principal harm is disclosure of sensitive information about people or data sets through the model itself.

Srida IT description in its own words. Refer to the licensed standard for ISO/IEC 27090's description, detection and mitigation guidance.

UNDERSTANDING THE COMPROMISE

B Description

Models tend to behave slightly differently on the data they were trained on than on data they have never seen. A classifier may be more confident on training records; a generative model may assign them higher likelihood or reproduce them more fluently. Membership inference exploits that difference. The attacker holds a candidate record, such as a person's attributes, a medical image, a document or a text passage, and queries the model to decide whether the record was a member of the training set. The model's weights are not stolen and no data is copied out directly; the leak is the single bit of membership, which can itself be highly sensitive. Knowing that a person's record trained a model of patients with a particular diagnosis, customers in arrears or members of a specific programme reveals that the person has that attribute.

Attacks vary in the access they require and in their strength. Simple **threshold attacks** compare the model's confidence or loss on the candidate with a calibrated threshold. **Shadow-model** attacks train substitute models on data from a similar distribution to learn what member and non-member outputs look like, then apply that learned distinction to the target. **Label-only** attacks work even where only a final decision is returned, by observing how stable the decision is under small changes to the input. White-box attacks use gradients or internal activations where the attacker has the model. For generative AI and large language models, membership inference extends to asking whether a specific document, code repository or copyrighted work was included in pre-training or fine-tuning, and it is closely related to memorisation and verbatim regurgitation of training data, which is addressed under sensitive model output (6.8).

Membership inference is a measurable form of privacy leakage, and the main factors that enable it, namely overfitting, small or unique data sets, outliers in the training data, and detailed outputs, are well understood. It is distinct from model inversion (6.6), which reconstructs representative features of a class or individual, and from model exfiltration (6.5), which copies model functionality or parameters. In practice the three share defences: limiting output detail, controlling query volume and reducing what the model has memorised. Federated learning does not by itself prevent membership inference, because shared updates can also leak membership; ISO/IEC 27090 covers federated machine learning in an informative annex.

C Attack Surface

Attribute	Detail
AI life-cycle stages	Design and development, Verification and validation, Deployment, Operation and monitoring, Retirement
AI assets targeted	Training data, Model parameters and weights, Model outputs, AI APIs and endpoints, Third-party models and AI services

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

Attribute	Detail
AI system types	Classical ML classifiers and risk scores trained on personal data (health, credit, insurance, HR), Medical imaging and genomics models, Generative AI and large language models fine-tuned on internal or customer data, Recommendation and personalisation models, Federated learning systems, Models shared with partners or released as open weights
Attacker access required	Black-box query access to scores, likelihoods or labels; possession of candidate records about the targeted individuals; for stronger attacks, white-box access to a released, shared or exfiltrated model, or participation in federated training.

D Resulting Failures and Harms

- Confidentiality and privacy breach: revealing that an individual's data was in a training set discloses the attribute that defines that set, which may constitute a personal data breach under applicable data protection law.
- Exposure of special-category data: membership of models trained on health, financial distress, criminal justice or similar data reveals sensitive attributes about identifiable people.
- Contractual and intellectual property exposure: a third party can show that licensed, confidential or copyrighted material was used in training, contrary to licence or consent terms.
- Loss of anonymisation claims: a model or data release described as non-personal may be shown to carry personal information, undermining the lawful basis for its sharing.
- Regulatory and reputational harm, including obligations to notify, retrain or withdraw the model.

E Attack Scenario

A hospital group trains a readmission-risk model on discharge records of patients treated in a specialist unit and exposes it through an API to partner clinics, returning a calibrated probability for each submission. A researcher at a partner organisation, holding demographic and clinical details of several local residents, submits their records and observes that some receive unusually confident predictions consistent with having been seen in training. Combining these results with knowledge of the unit's specialism, the researcher infers that those residents were treated there. The model had been trained for many epochs on a small cohort without regularisation, outputs were returned at full precision, and the API had no per-partner query budget or monitoring of repeated lookups. An authorised privacy evaluation before release (AIS-19) would have shown a membership advantage well above random guessing. Privacy-preserving training and data minimisation (AIS-03), coarsened outputs (AIS-08), access control and rate limits (AIS-06, AIS-09) and query monitoring (AIS-10) would together have reduced the leakage to an acceptable level.

F Likelihood and Impact Factors

Likelihood factors	Impact factors
• Degree of overfitting and memorisation: small training sets, many training epochs, large models relative to data and absence of regularisation increase leakage. • Detail of outputs: full-precision probabilities, logits or token likelihoods enable stronger attacks than labels alone. • Availability of candidate records to the attacker, which is high for partners, insiders, and data available from breaches or public sources. • Breadth of query access and whether the model or its weights have been shared, published or exfiltrated.	• Sensitivity of the attribute that defines membership of the training set. • Number of individuals in the training data and whether outliers or rare cases are present. • Legal obligations attached to the data, including data protection, health-information and contractual confidentiality. • Whether the model is distributed to third parties, which removes the organisation's ability to throttle or monitor queries.

DETECTION AND MITIGATION

G Detection

Indicators

- Queries that submit complete, realistic records that closely match training records, rather than the incomplete records typical of operational use.
- Repeated submission of the same or minimally varied record, consistent with label-only or stability-based attacks.

© 2026 Srida IT · Version 1.0 · Independent guidance, not an official ISO/IEC publication

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

- A single client submitting records about many different named or identifiable individuals outside its operational population.
- Query volumes or diversity from a client inconsistent with its business purpose.
- Pre-release privacy evaluations showing a member-versus-non-member advantage above the approved threshold.

Telemetry required

- Inference request logs with client identity, input fingerprint and output detail, protected as personal data themselves.
- Per-client query budgets, consumption and exception records.
- Privacy evaluation reports for each model release, including attack methods and measured advantage.
- Model cards or technical documentation recording training data scope, regularisation and privacy techniques.

Detection methods

- Pre-release membership inference evaluation using shadow-model or threshold attacks under authorisation, measured against an approved acceptance criterion.
- Query-pattern monitoring for repeated or record-like lookups and out-of-purpose populations (AIS-10).
- Fingerprint comparison of incoming queries against a protected hash set of training records, where proportionate and lawful.
- Monitoring of privacy metrics across releases and retraining (AIS-21).

H Mitigations

Preventive	Detective	Corrective
• Minimise personal data in training sets, de-identify where feasible and apply privacy-preserving training such as differential privacy, with the privacy budget recorded and approved (AIS-03). • Reduce overfitting through regularisation, early stopping and adequate data volume, and remove or treat unique outliers where lawful. • Return labels or coarse bands rather than full-precision scores or likelihoods unless the business need is documented (AIS-08). • Restrict model access to authorised clients with defined purposes, and protect model weights from disclosure (AIS-06, AIS-09). • Assess privacy risk in the data protection impact assessment and AI system impact assessment before release.	• Monitor query patterns for membership-probing behaviour and out-of-purpose lookups (AIS-10, AIS-20). • Re-run privacy evaluations after retraining or significant data changes (AIS-19, AIS-21).	• Treat confirmed leakage as a potential personal data breach under the incident process, including assessment of notification obligations (AIS-22). • Retrain with stronger privacy protection or remove affected records, and securely retire leaking model versions and their copies (AIS-24).

SAMPLE

I Srida IT AI Security Controls

- **AIS-03** Privacy-preserving training and data minimisation
- **AIS-06** Model access control and confidentiality
- **AIS-08** Output information minimisation
- **AIS-09** AI API authentication, rate limiting and throttling
- **AIS-10** Query pattern monitoring and extraction detection
- **AIS-14** Sensitive output filtering and leakage prevention
- **AIS-19** AI security testing and red teaming
- **AIS-20** AI security logging and telemetry
- **AIS-21** Model and mitigation deterioration monitoring
- **AIS-22** AI security incident response and model rollback
- **AIS-23** Federated learning security
- **AIS-24** Secure retirement of AI models and data

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

J Assessment Testing Procedures

Recommended Srida IT assessment methodology; not prescribed by ISO/IEC 27090. Adversarial tests are performed only under written authorisation and agreed rules of engagement (see Chapter 19 and Appendix F).

Table 12-3. Assessment testing procedures, Threat 6.4

Test ID	Method	Testing procedure	Result
TP-T6.4-01	**EXAMINE**	Examine the threat model, data protection impact assessment and AI system impact assessment for models trained on personal or confidential data, and confirm that membership inference is assessed with reference to the sensitivity of the training population and the access granted to each client type.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A
TP-T6.4-02	**EXAMINE**	Examine model technical documentation for the sampled models to confirm that training data scope, minimisation decisions, de-identification steps, regularisation and any privacy-preserving technique (including the differential privacy budget where used) are recorded and approved by the model owner and the privacy function.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A
TP-T6.4-03	**ADVERSARIAL TEST**	Under written authorisation and agreed rules of engagement, and with the privacy function's approval for handling the training data, perform a membership inference evaluation against a non-production copy of the model using recognised shadow-model, threshold and, where outputs are label-only, label-only methods. Use a held-out non-member set drawn from the same distribution. Report the attack advantage or true-positive rate at a low false-positive rate against the approved acceptance criterion. Do not record identities of inferred members in the report.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A
TP-T6.4-04	**INSPECT**	Inspect API responses for each client tier and confirm the precision of returned scores or likelihoods, and that any return of full-precision outputs or logits is justified and approved.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A
TP-T6.4-05	**INSPECT**	Inspect access control and query budget configuration for the model endpoints, confirming that each client is authorised for a documented purpose, that budgets are enforced, and that model weights are not distributed without a recorded privacy risk decision.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A
TP-T6.4-06	**SAMPLE**	Select a sample of query-monitoring alerts or, if none, a sample of high-volume client sessions from the period (Srida IT recommended practice: 25 or all if fewer). Confirm each was reviewed for membership-probing patterns, and that the inference logs used are themselves access-restricted and retained only as long as necessary.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A
TP-T6.4-07	**EXAMINE**	Examine records of retraining events in the period and confirm that privacy evaluation was repeated where training data or architecture changed, and that results were compared with the previous release.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A
TP-T6.4-08	**INTERVIEW**	Interview the data protection officer and model owner to confirm that membership inference leakage is recognised as a potential personal data breach scenario in the incident process, and that model retirement procedures cover copies held by partners.	☐ In place ☐ In place w/ obs. ☐ Not in place ☐ N/A

K Interview Questions

Interviewee	Question
Data protection officer	Would it be a breach if someone could show that a named person's record trained this model? How have you assessed that risk?
Model owner	Why does this model return full-precision scores, and which clients actually need them?

SAMPLE

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT

Interviewee	Question
ML engineer	What did the last privacy evaluation measure, against what threshold, and what did you change as a result?
Partner or integration manager	Which external parties hold copies of this model or its weights, and under what terms?

L Expected Evidence

- DPIA and AI system impact assessment sections addressing membership inference.
- Model technical documentation recording minimisation, de-identification, regularisation and privacy techniques.
- Authorised privacy evaluation plans and reports with measured attack advantage and acceptance criteria.
- API configuration showing output precision, authentication and query budgets per client tier.
- Query-monitoring rules, alert samples and review records.
- Register of external parties holding model copies, with contractual terms and retirement records.

M Evaluation Criteria

Effective

- Models trained on sensitive data have a documented privacy evaluation showing membership advantage within an approved criterion for the current release.
- Outputs are limited to the precision the use case needs, and access is purpose-bound, throttled and monitored.
- Leakage is recognised as a breach scenario with a defined response, including retraining and retirement.

Partially effective

- Privacy evaluation was performed once but not repeated after retraining.
- Minimisation and de-identification are applied but full-precision outputs remain available to all clients.
- Query monitoring exists for volume only, without patterns characteristic of membership probing.

Ineffective

- Membership inference has not been considered for models trained on special-category or confidential data.
- Models or weights trained on personal data are shared externally without privacy evaluation or contractual control.
- No query limits or monitoring apply to an externally accessible model trained on personal data.

N Common Gaps and Findings

- Finding: The readmission-risk model, trained on records of a specialist unit, returns six-decimal probabilities to all partner clinics; no business need for this precision is documented.
- Finding: No membership inference or other privacy evaluation has been performed on any of the three models trained on customer financial data.
- Finding: The DPIA for the customer-support LLM fine-tuning project does not consider memorisation or membership inference of support transcripts.
- Finding: Model weights trained on employee HR data were shared with a vendor for optimisation without a privacy risk decision or retirement clause.
- Finding: Inference logs containing full input records are retained indefinitely and accessible to the whole data science team.

O Monitoring Metrics

- Percentage of models trained on personal data with a current privacy evaluation within acceptance criteria.
- Measured membership inference advantage per model release, trended.
- Percentage of model endpoints returning output at higher precision than documented need.
- Number of external parties holding model copies without a recorded privacy decision.

P Related Standards and Frameworks

- ISO/IEC 27090 Clause 6 (6.4 Membership inference); related compromise types 6.5 Model exfiltration, 6.6 Model inversion and 6.8 Sensitive model output; informative annex on federated machine learning.
- ISO/IEC 27001:2022 Annex A: A.5.34 Privacy and protection of PII; A.8.11 Data masking; A.8.12 Data leakage prevention; A.5.15 Access control; A.8.16 Monitoring activities.
- ISO/IEC 42001:2023 Annex A: A.5.2 AI system impact assessment process; A.7.2 Data for development and enhancement of AI system; A.7.4 Quality of data for AI systems; A.6.2.4 AI system verification and validation.

© 2026 Srida IT · Version 1.0 · Independent guidance, not an official ISO/IEC publication

Free sample pages · The complete manual is available at sridait.com/manuals/rtp-iso27090 · © 2026 Srida IT