AI SECURITY ENGINEERING HANDBOOK · 2026

Chapter 07 · Data Exposure and Privacy

Standalone study module for LMS delivery and required reading.

FORMAT	USE	SCOPE	AUDIENCE
Standalone PDF	Study module	Single chapter	Learners

CHAPTER 07

Data Exposure and Privacy

HANDBOOK STUDY COMPANION: STUDY FRAME

Use this chapter to build vocabulary, judgment, and role-readiness. Pair it with the Field Guide when you need applied actions, checklists, and control execution.

STUDY FOCUS

STUDY FOCUS	WHY IT MATTERS
Prompt, embedding, log, memory, output, and vendor data flows, with privacy controls and evidence expectations.	AI features can move sensitive data into new contexts faster than privacy and security processes detect.

Study Outcomes

- Identify sensitive data paths in AI workflows.
- Explain minimization, retention, logging, and deletion evidence.
- Connect privacy obligations to engineering controls.

DOMAIN MAPPING

RELATED AIPSA DOMAINS	APPLIED NEXT STEP	WORKBENCH INSTRUMENTS	RELATED SERVICES
Privacy and Data Protection in AI Systems	Privacy and data protection	Runtime Proxy, AI Control Crosswalk	AI Product Security Assessment

CERTIFICATION AND ASSESSMENT BOUNDARY

This chapter supports training, diagnostic preparation, scorecards, interviews, and role-readiness evaluation. It does not guarantee credential outcomes.

The privacy challenge in AI systems is that data moves in directions that product teams do not track and security teams do not audit. A customer support message becomes a fine-tuning example, then a retrieval embedding, then an eval fixture, then an inference-time retrieval result — each transformation creating a new derived representation with its own retention lifecycle and its own deletion requirement. Traditional privacy programs were designed to track records in databases. AI systems create derived representations in vector indexes, model weights, prompt logs, and annotation queues that those programs were never designed to govern.

Traditional privacy programs were designed to track records in databases. AI systems create derived representations in vector indexes, model weights, prompt logs, and annotation queues that those programs were never designed to govern.

HANDBOOK

CORE CONCEPTS

SOURCE-TO-DERIVATIVE LINEAGE

Every AI-specific transformation of personal data — from source document to chunk to embedding to index entry, from customer interaction to prompt log to fine-tuning example — creates a derived representation with independent privacy obligations. Lineage tracking maps each derived artifact back to its source record so that deletion, reclassification, or consent withdrawal can propagate correctly. Without lineage, the organization cannot honor erasure requests with confidence, cannot scope a privacy incident accurately, and cannot demonstrate compliance to a regulator or auditor.

DELETION PROPAGATION TO AI ARTIFACTS

Deleting the source record is the first step, not the complete operation. The organization must also identify and act on: embedding records in the vector index, cached responses that used the source data, prompt logs that included the source content, fine-tuning dataset entries, eval fixtures, and annotation records. Each artifact type has different deletion mechanics. Vector index deletion requires either record-level deletion with confirmed propagation or index rebuild from a cleaned source. Model artifact deletion may not be technically feasible, requiring instead exclusion from future training runs and disclosure of the limitation.

PURPOSE LIMITATION FOR AI PROCESSING

Data collected for one purpose cannot be freely reused for AI processing. Customer support conversations collected for service delivery may not be used for model training without a separate legal basis and disclosure. Product interaction data collected for analytics may not be used for fine-tuning without consent. Purpose limitation requires review at three points: when a dataset is created or assembled for AI use, when a model or embedding is trained or fine-tuned, and when a vendor receives data for AI processing. Each use case needs its own legal basis analysis.

PROMPT LOG PRIVACY DESIGN

Prompt logs may contain personal data that users entered in their queries, personal data about third parties referenced in prompts, credentials inadvertently included in context, confidential business information, and regulated health or financial data. A prompt logging policy defines logging scope by sensitivity tier, redaction requirements for known sensitive fields, access controls for each tier, retention periods calibrated to sensitivity, and break-glass access procedures for high-sensitivity logs. The policy must balance investigation requirements with minimization obligations.

VENDOR AI PROCESSING SCOPE

Model providers, embedding services, annotation vendors, and AI quality platforms all create processing relationships with distinct privacy obligations. Each vendor may retain prompt and response data for defined periods, use it for model improvement unless opted out, route it through sub-processors, and apply different security standards to it than the primary contract suggests.

The organization's privacy notice and data processing agreements must accurately reflect every vendor that processes personal data through AI workflows — including vendors added through product experimentation that bypassed procurement review.

THE PRACTITIONER'S CHALLENGE

The political challenge is that privacy review is perceived as a launch blocker rather than a design input. Teams building AI features frequently discover privacy obligations after the system is already working, when changing the data flow requires engineering rework. The practitioner must establish privacy review as a design-time requirement — a question about data lineage and vendor processing that is answered before the architecture is fixed — rather than a launch-time approval.

The structural challenge is that AI privacy crosses multiple team boundaries. Engineering owns the AI features and data pipelines. ML platform owns training and fine-tuning workflows. Legal and privacy own the data processing agreements and legal basis analysis. Procurement owns vendor contracts. Security owns logging policy and access control. Each team holds a piece of the privacy picture. A complete AI privacy program requires explicit ownership of the cross-team data lineage map and a defined handoff between engineering data flows and legal/compliance obligations.

The technical challenge is that some AI privacy failures are not solvable through conventional controls. A model trained on personal data may memorize and reproduce that data during inference. An embedding derived from personal data carries semantic information that can be used for re-identification. These issues require architectural decisions — what data to include in training, how to test for memorization, whether the use case justifies the privacy risk — not just operational controls. The practitioner must identify when a privacy problem requires an architectural response.

HOW TO APPROACH IT

- Start with a data lineage map for each AI feature or system. Trace every path that personal data takes from first entry through AI-specific transformations: ingestion to embedding to index, customer interaction to prompt log to retrieval result, conversation record to fine-tuning example to model artifact. For each derived representation, document its storage location, retention period, access controls, deletion mechanics, and the lineage record that connects it to the source.

- Specify deletion propagation requirements for each AI artifact type. For vector index entries, define the maximum acceptable propagation latency and the immediate invalidation procedure. For prompt logs, define the retention tier and automatic expiration. For fine-tuning datasets, define the exclusion process when a subject requests deletion and document the limitation that model artifacts cannot be retroactively cleaned. For vendor records, define the deletion request process and the contractual timeline for confirmation.

- Write a prompt logging policy that defines sensitivity tiers before deployment. The policy should specify: what can be logged as metadata only, what requires redaction before logging, what can be logged in full under restricted access, who can access each tier, what the retention period is for each tier, and what the break-glass access procedure is for high-sensitivity logs. The policy should be reviewed by privacy counsel and engineering together, not written by either in isolation.

- Review every AI vendor relationship for data processing scope. For each vendor that receives personal data — model provider, embedding service, annotation platform, AI quality vendor — review the data processing agreement for: retention period, training-on-input default and opt-out configuration, sub-processor list, geographic routing, breach notification timeline, and deletion request process. Verify that the API configuration matches the contracted terms. Document the processing scope in the organization's privacy notice.

- Build privacy testing into the development workflow. For vector indexes, run deletion propagation tests before launch: ingest records, delete source records, verify chunk disappearance with timing. For retrieval systems, test that low-privilege queries do not return personal data belonging to other users. For prompt logs, verify redaction rules are working as designed. These tests confirm that the privacy controls are implemented correctly, not just specified.

OUTPUTS AND DELIVERABLES

- The design artifacts are the **AI data lineage map**, **personal data inventory for AI systems**, and **purpose limitation analysis**. The lineage map shows every transformation of personal data through AI-specific workflows with retention periods and deletion mechanics for each derived artifact. The personal data inventory for AI systems identifies each data category, its AI processing use cases, the legal basis for each, and the vendor processing relationships. The purpose limitation analysis documents the legal basis review for each AI processing use case.

- The operational artifacts are the **prompt logging policy**, **deletion propagation specification**, and **AI vendor privacy assessment template**. The logging policy defines sensitivity tiers, redaction rules, access controls, and retention periods. The deletion propagation specification defines requirements and test procedures for each AI artifact type. The vendor assessment template covers retention terms, training opt-out configuration, sub-processors, geographic routing, and deletion procedures.

- The evidence artifacts are the **deletion propagation test records**, **privacy notice accuracy review**, and **data processing agreement compliance checklist**. Deletion tests confirm that propagation mechanics work correctly. The privacy notice review confirms that all AI processing is accurately disclosed. The DPA checklist confirms that vendor contracts match actual API configuration and sub-processor scope.

COMMON FAILURE MODES

- **Source-Record-Only Deletion:** The team honors deletion requests by deleting the source record and considers the obligation satisfied. Embeddings, prompt logs, fine-tuning examples, and cached responses derived from the source data persist. Fix: build source-to-derivative lineage tracking and define deletion mechanics for each artifact type before handling the first deletion request.

- **Undisclosed Vendor Processing:** An AI vendor added through product experimentation processes personal data without appearing in the privacy notice or data processing agreement. The processing is discovered during a customer question or regulatory inquiry. Fix: require privacy review of every new AI vendor before API key provisioning; connect AI vendor inventory to the privacy notice update process.

- **Prompt Log Sprawl:** Engineering enables comprehensive logging for debugging without privacy classification. Over time, logs accumulate sensitive personal data from customer queries with broad engineering access and undefined retention. Fix: write the prompt logging policy before enabling logging; treat prompt logs as a sensitive data category from the first line of code.

- **Purpose Creep in Training:** Customer interaction data collected for service delivery gets included in a fine-tuning dataset without legal basis review. The model is trained and deployed. Fix: require purpose limitation analysis as a gate for any dataset assembled for AI training or fine-tuning; make this review a prerequisite for ML platform access to production data exports.

IMPLEMENTATION CHECKLIST

- [ ] Build a data lineage map for each AI feature covering every AI-specific data transformation.
- [ ] Define deletion mechanics for each AI artifact type: vector records, prompt logs, fine-tuning examples, cached responses.
- [ ] Write a prompt logging policy before enabling any logging, with sensitivity tiers and access controls.
- [ ] Review every AI vendor data processing agreement for retention, training opt-out, sub-processors, and deletion procedures.
- [ ] Run deletion propagation tests before launch for every system with vector indexes.
- [ ] Require purpose limitation analysis as a gate for datasets used in AI training or fine-tuning.
- [ ] Verify that the privacy notice accurately reflects every AI processing pathway and vendor relationship.
- [ ] Test retrieval systems for personal data leakage across users and tenants.

RELATED READING

- Handbook chapters: Chapter 5 (RAG Authorization) for retrieval-time data access controls that prevent unauthorized personal data access; Chapter 10 (Logging and Telemetry) for prompt log design; Chapter 14 (Governance Evidence and Customer Trust) for privacy evidence requirements.
- Field Guide: Privacy and Data Protection in AI Systems for data-flow review, deletion propagation checks, and provider processing evidence.