# Day 14

Generative AI Security

## AI Security Policies — Data Usage Policies:

A Data Usage Policy defines how data can be used safely and legally for AI systems—especially for training and inference. Because AI systems depend heavily on data, bad data practices = high security, privacy, and legal risk.

**Rules for Collecting, Storing, Processing, and Sharing AI Data:**

Data Collection: collect data from approved sources only, verify legal rights to use the data, prohibit scraping, avoid collecting personal data.

Data Storage: store datasets in secure repositories, encrypt data at rest, restrict access based on the roles.

Data processing: track all transformations, document preprocessing steps, avoid mixing datasets without approval.

Data Sharing: data shared externally requires approval, use data-sharing agreements, never share personal data.

**Data Classification:**

Data is classified to apply the right level of protection.

Common AI Data Classifications

- Public – safe to share
- Internal – limited internal use
- Confidential – business-sensitive
- Sensitive / Restricted – PII, financial, health data

Why classification matters

- determines encryption level
- determines access control
- defines where data can be used (training vs inference)

**Privacy and PII Handling:**

PII Handling Rules

- identify PII before training
- mask, anonymize, or pseudonymize PII
- restrict use of sensitive attributes (race, health, biometrics)
- enforce consent where required

Privacy Techniques

- anonymization (irreversible)
- pseudonymization (reversible with keys)
- differential privacy
- federated learning (where possible)

**Data Minimization Principles**

What is data minimization? Use only the minimum data required to achieve the AI purpose.

Rules

- do not collect “just in case” data
- avoid long-term storage of unused data
- delete data after purpose is achieved
- regularly review dataset necessity

Example: If age range is enough → do not store full date of birth.

**Usage Restrictions for Proprietary or Third-Party Datasets:**

Proprietary Data

- cannot be used outside approved projects
- cannot be shared with vendors or LLMs
- requires business owner approval

Third-Party Datasets

- must follow licensing terms
- usage must match allowed purposes
- training vs inference use must be checked
- redistribution is usually forbidden

**Synthetic Data Usage Rules:**

What is synthetic data? Artificially generated data that imitates real data.

**Rules**

- synthetic data must not re-identify real individuals
- validate synthetic data quality
- test for bias amplification
- document how synthetic data was generated

Synthetic data is helpful, but not risk-free.

**Auditing & Logging Data Access and Usage**

What must be logged

- who accessed the data
- when it was accessed
- purpose of access

- which system or model used it
- what changes were made

Why auditing matters

- supports investigations
- provides compliance evidence
- detects misuse or insider threats

Logs must be protected and reviewed regularly.

Basically, **SUMMARY:**

- Data usage policies control how AI data is collected, used, stored, and shared
- Data must be classified and protected accordingly
- PII must be anonymized or minimized
- Use only necessary data
- Proprietary and third-party data have strict restrictions
- Synthetic data must be validated
- All data access must be logged and audited

## AI Security Policies — Model Access Policy:

A Model Access Policy defines who can access AI models, what they are allowed to do, and under what conditions. Its goal is to prevent misuse, leakage, tampering, and unauthorized use of AI models.

**Access Control Methods for AI Models:** Access control decides who gets access and how much.

1. RBAC (Role-Based Access Control): Access is given based on job role.
2. ABAC (Attribute-Based Access Control): Access depends on attributes, not just roles.
3. Privileged Access Management (PAM): Used for high-risk actions.

**Authentication & Authorization for Model Use:**

Authentication (Who are you?)

Authorization (What can you do?)

Defines:

- which model can be accessed
- allowed actions (read, infer, train, deploy)
- data allowed during inference

**Access Approval Workflows:** Access must never be granted automatically.

Typical Workflow

1. Access request submitted
2. Business justification provided
3. Manager approval
4. Security approval
5. Time-bound access granted

6. Access reviewed periodically

High-risk models require governance board approval.

**Segmentation of Access (Training, Evaluation, Production):** Different environments = different risks. Segmentation prevents accidental damage and data leaks.

Training Models

- accessible only to ML teams
- use restricted datasets
- no external access

Evaluation / Testing Models

- used by QA, red team, compliance
- limited data access
- monitored closely

Production Models

- strictest access
- inference only
- no weight downloads

**Versioning and Rollback Permissions:** All rollback actions must be logged and approved.

Model Versioning:

- every model version must be tracked
- changes must be documented
- access to old versions is restricted

Rollback Permissions

Only authorized roles can:

- roll back models
- disable a deployed model
- switch traffic to older versions

**Handling Third-Party Access or API Integration Securely:** Third-party access is high risk.

Security Rules:

- use separate API keys or service accounts
- enforce rate limits
- restrict accessible models

Contract & Legal Controls

- data usage restrictions
- model IP ownership clarified
- audit rights defined
- breach notification clauses

Basically, **SUMMARY:**

- Model access policy controls who can use AI models and how
- Use RBAC, ABAC, and PAM for layered security
- Strong authentication + authorization is mandatory
- Access requires formal approval workflows
- Training, testing, and production access must be separate
- Version changes and rollbacks are restricted
- Third-party access must be limited, monitored, and contractually controlled

--The End--