Day 15

Generative AI Security

AI Security Policies — Model Training Policies:

A Model Training Policy defines how AI models are trained safely, ethically, and securely so that training does not introduce bias, security risks, or compliance violations. It ensures that only trusted data, approved environments, and controlled processes are used to train AI models.

Secure Model Training Environment Requirements: The training environment is where models learn — if it is insecure, the model itself becomes unsafe.

Key Requirements

- Training must happen only in approved environments (cloud accounts / on-prem clusters)
- No personal laptops or unapproved GPUs
- Network isolation (private VPC / subnet)
- No direct internet access unless approved
- Secrets (API keys, tokens) stored in vaults
- Access limited to ML engineers with approval

Why This Matters

If attackers access the training environment, they can:

- poison data
- steal model weights
- inject backdoors

Approved Datasets and Pre-Processing Rules:

Approved Datasets

Only datasets that are:

- reviewed by data & compliance teams
- documented with source, license, and purpose
- privacy-cleared (PII handled)

Examples:

- internal business data (approved)
- licensed third-party datasets
- approved synthetic data

Pre-Processing Rules

Rules applied before training:

- remove PII where not required
- normalize and clean data
- detect duplicates and outliers
- label verification
- log every transformation

Example:
If customer data is used, names and phone numbers must be anonymized before training.

Bias Mitigation Strategies: Bias can be introduced during training if data is unfair or unbalanced.

Common Strategies

- Dataset balance checks (gender, geography, language)
- Removing proxy attributes (e.g., ZIP code for race)
- Re-sampling underrepresented groups
- Fairness constraints during training
- Separate evaluation for sensitive attributes

Policy Requirement: Bias testing is mandatory before deployment for high-impact models.

Security Checks During Training: Poisoning Prevention: Attackers may inject malicious data to manipulate behavior.

Controls include:

- dataset integrity checks (hashing)
- anomaly detection in training data
- restricted data upload permissions
- manual review for critical datasets

Adversarial Robustness**:** Models must be trained to resist attacks.

Techniques:

- adversarial training
- noise injection
- regularization
- testing against malicious inputs

Why This Matters

Without these checks:

- models can be manipulated
- outputs can be unsafe
- attackers can create hidden triggers

Logging and Monitoring Training Processes: Everything during training must be logged and monitored.

What to Log

- who started training
- dataset versions used
- code version (Git commit)
- hyperparameters
- training duration

Monitoring

- abnormal training time
- sudden accuracy jumps or drops
- unexpected data size changes
- unauthorized access attempts

- compute resources
- warnings or anomalies

Audit Trail of Model Training Metadata: An audit trail is proof of how the model was trained.

Metadata to Store

- model version ID
- training date and time
- dataset names and hashes
- feature sets
- hyperparameters
- model weights checksum
- training environment details
- approvals and sign-offs

Why Audit Trails Matter

- regulatory compliance
- incident investigations
- reproducibility
- rollback and retraining

Example: If a model causes harm, the organization must prove what data and settings were used.

Basically,

- Model training must happen in secure, approved environments
- Only approved and documented datasets can be used
- Data must be cleaned, privacy-safe, and bias-checked
- Security checks prevent poisoning and adversarial attacks
- Training activities must be fully logged and monitored
- Complete audit trails ensure compliance, accountability, and trust

AI Security Policies — Third-Party AI Tools Usage Policy:

A Third-Party AI Tools Usage Policy defines how an organization can safely use external AI tools, platforms, or APIs (like LLM APIs, AI SaaS tools, or cloud AI services) without exposing data, violating laws, or losing control.

Examples of third-party AI tools:

- External LLM APIs
- AI SaaS platforms
- Vendor-hosted vision, speech, or NLP models
- Plug-in based AI services

Security and Compliance Considerations: Before using any external AI tool, the organization must ensure it is secure and legally compliant.

Key Security Checks

- Strong authentication (API keys, OAuth)
- Encryption in transit and at rest
- Access logging and audit trails
- Isolation of customer data
- Secure model update process

Compliance Checks

- GDPR / DPDP / sector-specific laws
- Data residency requirements
- Right to delete data
- Incident disclosure timelines

Example: A public LLM API cannot be used if it stores prompts permanently without consent.

Vendor Risk Assessment Before Procurement: Every AI vendor must go through a formal risk assessment. High-risk vendors need governance board approval.

Assessment Areas

- Vendor security posture (ISO, SOC 2, etc.)
- Data handling and retention policies
- Model training practices
- History of breaches or incidents
- Dependency and lock-in risks

Risk Classification

- Low-risk: generic tools, no sensitive data
- Medium-risk: internal business data
- High-risk: PII, regulated or critical systems

Licensing and Intellectual Property Considerations: AI tools often have complex licensing rules.

Key Questions

- Who owns the generated outputs?
- Can outputs be used commercially?
- Is training on customer data allowed?
- Are there usage limits or restrictions?

Policy Rules

- No tool may be used without license review
- Training internal models using vendor output may be restricted
- Generated content must not violate copyright

Example: Some tools allow usage for internal purposes but restrict resale or redistribution.

Data Sharing & Privacy Rules: Third-party tools must follow strict data rules.

Allowed Data

- Public data
- Anonymized or synthetic data
- Approved internal data (case-by-case)

Prohibited Data

- Personal data without consent
- Credentials, secrets, keys
- Regulated or classified information

Privacy Controls

- Data minimization
- Prompt redaction
- PII masking before submission
- Opt-out from vendor training (where possible)

Monitoring and Restricting Output Usage: Third-party AI outputs can also be risky.

Risks

- hallucinations
- biased or harmful content
- copyright infringement
- unsafe recommendations

Policy Controls

- Human review for high-impact outputs
- Output filtering and moderation
- Prohibiting blind trust in outputs
- Usage logging and traceability

Example: AI-generated legal or medical advice must always be reviewed by a qualified human.

Periodic Review of Third-Party Tool Security: Vendor risk is not one-time.

Ongoing Reviews Include

- annual security re-assessment
- license and ToS changes
- new data handling practices
- model updates or architecture changes
- incident or breach history updates

Actions

- renew approval
- restrict usage
- require compensating controls
- decommission the tool

Basically,

- Third-party AI tools can introduce security, privacy, and legal risks
- Vendors must pass security and compliance checks
- Licensing and IP rights must be clearly understood
- Only approved data can be shared with external tools
- Outputs must be monitored and reviewed
- Vendor security must be reviewed regularly

--The End--