Day 17

AI Security Policies — Security Controls Across AI Model Lifecycle:

AI security is not a single step. Security controls must be applied continuously across the entire AI lifecycle — from data collection to monitoring in production. Think of it like multiple safety gates, not one lock.

Integrating Security Controls Across Each Lifecycle Stage:

A. Data Collection & Ingestion

Risks

- poisoned data
- privacy violations
- illegal data usage

Security Controls

- approved data sources only
- data integrity checks (hashing)
- PII detection and masking
- licensing verification
- access controls on data uploads

Example:
A dataset is rejected if its source or license is unknown.

B. Model Training

Risks

- backdoored models
- data leakage
- unauthorized training changes

Security Controls

- isolated training environments
- restricted access
- dataset version locking
- adversarial training
- bias and robustness checks
- full training logs

C. Model Evaluation & Validation

Risks

- unsafe behavior going unnoticed
- hidden vulnerabilities

Security Controls

- adversarial testing
- red teaming
- prompt injection tests
- bias and fairness evaluation
- stress testing edge cases
- independent validation teams

Example:
A model is blocked from deployment if jailbreak success rate is high.

D. Model Deployment

Risks

- unauthorized access
- misconfiguration
- insecure APIs

Security Controls

- access control (RBAC/ABAC)
- API authentication
- rate limiting
- environment segmentation
- approval workflows
- rollback readiness

E. Monitoring & Operations

Risks

- model drift
- abuse and misuse
- delayed incident detection

Security Controls

- output monitoring
- prompt logging
- anomaly detection
- drift detection
- alerting on policy violations
- SOC integration

Automated Security Checks and Alerts: Automation reduces human error and speeds up response.

Automated Checks

- data integrity scans
- prompt injection detection
- output safety classifiers
- access violation alerts
- drift threshold alerts

Alerts

- sent to SOC and AI teams
- severity-based (low, medium, high)
- linked to runbooks

Example: Repeated jailbreak attempts trigger an automatic alert.

Red-Teaming and Adversarial Testing

What is AI Red Teaming? Deliberately trying to break or misuse the model.

Activities

- prompt injection attempts
- data extraction attacks
- bias exploitation
- unsafe instruction testing
- supply chain manipulation tests

When to Red Team

- before deployment
- after major updates
- periodically for high-risk models

Incident Detection and Response Planning:

AI-Specific Incident Detection

- abnormal outputs
- sudden bias increase
- data leakage signs
- unauthorized access

Response Planning

- isolate affected model
- block unsafe prompts
- revoke access
- roll back to safe version
- notify stakeholders
- legal and compliance reporting

Response must be predefined and rehearsed.

Continuous Improvement and Feedback Loops: AI security must evolve continuously.

Feedback Sources

- incidents and near misses
- monitoring insights
- red team results
- user feedback
- regulatory changes

Improvements

- update controls
- retrain models
- improve filters
- strengthen policies
- retrain teams

This creates a security learning loop.

Summary:

- AI security controls must exist at every lifecycle stage
- Automation helps detect issues early
- Red teaming finds weaknesses before attackers do
- Incident response must be planned in advance
- Continuous feedback improves long-term security

AI Security Policies — Role of Risk Assessments in Policy Development:

Risk assessments are the foundation of good AI security policies. They help organizations decide what rules are needed, how strict they should be, and where controls must be applied. Without risk assessment, policies become either too weak or unnecessarily strict.

Performing AI Risk Assessments Before Policy Creation:

Why Do Risk Assessment First? Different AI systems have different risk levels.

Example:

- Internal chatbot → low risk
- AI approving loans → high risk
- AI diagnosing diseases → very high risk

Policies must be based on real risks, not assumptions.

What Is Assessed

- type of data used
- model behavior and autonomy
- impact of wrong decisions

- security exposure
- legal and ethical risks

This assessment happens before writing policy rules.

Identifying High-Risk AI Use Cases:

What Makes an AI Use Case High Risk

- handles personal or sensitive data
- affects people's rights or finances
- operates without human oversight
- makes irreversible decisions
- exposed to external users

Examples:

- recruitment screening AI
- credit scoring models
- healthcare AI
- public-facing GenAI systems

High-risk use cases need stricter policies.

Mapping Risks to Policy Rules and Controls: Risk assessments directly shape policy content.

Risk → Policy Mapping Examples

Identified Risk	**Policy Rule**
Data leakage	No PII in prompts
Bias	Mandatory bias testing
Prompt injection	Input validation & filters
Model misuse	Access controls & logging
Regulatory breach	Compliance approval required

This ensures every policy rule has a clear reason.

Periodic Risk Reassessment and Policy Updates: AI risks change over time.

Reasons:

- new threats (jailbreaks, poisoning)
- model updates
- new regulations
- new use cases
- incident learnings

Policy Requirement

- risk reassessment at regular intervals

- immediate reassessment after incidents
- policy updates approved by governance board

Policies must be living documents, not static PDFs.

Balancing Business Needs vs. Security Requirements:

The Reality

- Business wants speed and innovation
- Security wants control and safety

Risk assessment helps balance both.

Risk-Based Decisions

- low-risk use case → lighter controls
- high-risk use case → strong controls
- acceptable risk → documented exception
- unacceptable risk → use case blocked

Example: A marketing chatbot may not need HITL, but a loan approval AI must have it.

Summary:

- Risk assessments guide what policies are needed
- High-risk AI systems require stricter rules
- Every policy rule should map to a real risk
- Risks and policies must be reviewed regularly
- Risk assessment helps balance innovation and security

AI Security Policies — Role of Risk Assessments in Policy Development:

Internal standards define how AI can and cannot be used inside an organization. They convert high-level AI policies into clear, repeatable, day-to-day rules that teams can follow. Think of standards as "how we do AI safely here".

Defining "Acceptable" vs. "Prohibited" AI Use Cases

Acceptable AI Use Cases**:** These are allowed with defined controls.

Examples:

- internal knowledge search chatbot
- customer support assistant (with filters)
- fraud detection with human review
- document summarization for internal use

Conditions:

- approved data sources
- human oversight where required
- compliance checks completed

Prohibited AI Use Cases

These are not allowed under any circumstances.

Examples:

- using AI for mass surveillance
- social scoring of individuals
- autonomous legal or medical decisions
- training models on stolen or scraped private data
- generating malware or deepfakes

Prohibited use cases are blocked at policy level.

Standard Procedures for Model Deployment Approvals: Standards ensure no model reaches production without checks.

Typical Approval Flow

1. Use case definition submitted
2. Risk assessment performed
3. Security and privacy review
4. Bias and safety testing
5. Governance board approval (if high-risk)
6. Production deployment approval

Each step has clear owners and sign-offs.

Standard Procedures for Model Deployment Approvals: Standards ensure no model reaches production without checks.

Typical Approval Flow

1. Use case definition submitted
2. Risk assessment performed
3. Security and privacy review
4. Bias and safety testing
5. Governance board approval (if high-risk)
6. Production deployment approval

Each step has clear owners and sign-offs.

Documentation Templates for Policy Compliance: Standards require consistent documentation.

Common Templates

- AI use case proposal
- Risk assessment form
- Data source and lineage document
- Model card
- Security testing report
- Deployment approval record

Templates reduce confusion and ensure audit readiness.

Collaboration Rules

- early security involvement (not last-minute)
- shared dashboards and tickets
- regular review meetings
- clear RACI responsibilities

Example: Security reviews the architecture before model training starts.

Metrics and KPIs for Monitoring Policy Adherence: You cannot manage what you don't measure.

Example AI Governance Metrics

- % of AI use cases approved vs rejected
- % of models with completed risk assessments
- number of policy violations
- time to remediate AI security issues
- number of unsafe output incidents
- audit findings related to AI

Summary:

- Internal standards define how AI is used safely
- Acceptable and prohibited use cases must be clearly listed
- Model deployments follow standard approval workflows
- Documentation templates ensure consistency and audits
- Cross-team collaboration prevents silos
- KPIs help track policy compliance and effectiveness

--The End--