Data Governance and Security in AI

Security & Privacy in AI

Learning Objective

By the end of this lesson, learners will be able to **describe AI security threats, privacy challenges, and strategies to protect AI systems**.

Overview

AI security and privacy concerns are critical in ensuring **trustworthy AI deployments**. Organizations must safeguard AI models against cyber threats, ensure data privacy compliance, and implement best practices for securing AI pipelines.

1. AI Security Threats & Risk Mitigation

Common AI Security Threats

1. **Adversarial Attacks**
    - Attackers manipulate AI inputs to deceive models.
    - Example: **A slight modification to an image causes misclassification in vision AI.**
2. **Data Poisoning Attacks**
    - Malicious data inserted into training datasets to alter model behavior.
    - Example: **Corrupting training data to manipulate fraud detection AI.**
3. **Unauthorized Model Access**
    - Exposing API endpoints or model weights can lead to unauthorized use.
    - **Solution:** Implement **rate-limiting, authentication, and encryption.**

Risk Mitigation Strategies

- **Secure Model Deployment:**
    - Apply **encryption** for model storage and API communication.
    - Use **access controls** to restrict unauthorized users.
- **Continuous Monitoring & Auditing:**
    - Implement **log analysis and anomaly detection** to flag security threats.
    - Use **version control** to track and validate model updates.

2. Privacy Considerations in AI

Key Privacy Challenges

1. **Personally Identifiable Information (PII) Exposure**
    - AI models often process sensitive data that must be anonymized.
    - **Solution:** Use **differential privacy techniques** to protect user data.
2. **Regulatory Compliance (GDPR, CCPA, etc.)**
    - AI models must adhere to global privacy regulations.
    - **Solution:** Implement **automated compliance checks** in data pipelines.
3. **Model Inference Attacks**
    - Attackers extract training data by querying the model.
    - **Solution:** Use **privacy-preserving AI techniques** (e.g., federated learning).

Best Practices for Privacy Protection

- **Data Anonymization & Masking**
- **Consent Management & Transparent Data Policies**
- **Privacy-by-Design Implementation**

Hands-On Activity: Security & Privacy Risk Assessment

Scenario: A client in the **healthcare sector** is deploying an AI model to analyze patient data for predictive diagnostics. The model processes **sensitive health records**, raising security and privacy concerns.

Task:

1. **Identify three key security and privacy risks** that could arise.
2. **Propose mitigation strategies** using security frameworks and privacy best practices.
3. **Develop a client action plan** detailing recommendations for safe AI deployment.

Key Takeaways

- **AI security risks** include adversarial attacks, data poisoning, and unauthorized access.
- **Privacy challenges** include PII exposure, compliance issues, and inference attacks.
- **Mitigation strategies** include encryption, monitoring, anonymization, and compliance frameworks.
- **Organizations must integrate security and privacy measures** into AI lifecycle management.

< Previous

Next >