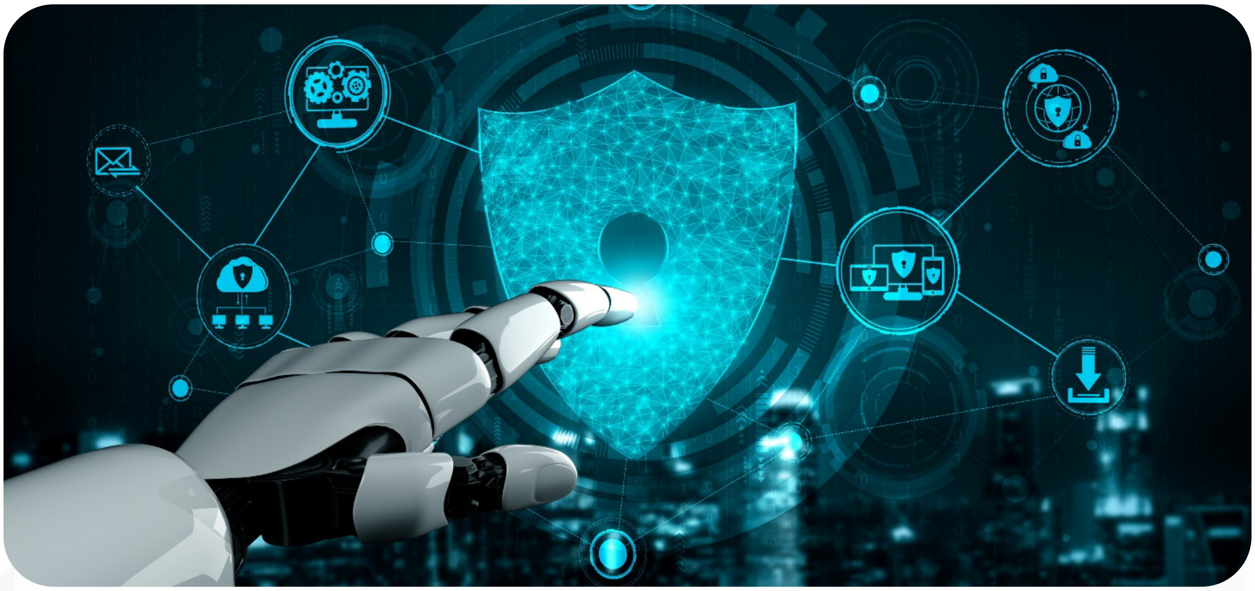
SECURING AI MODELS

FOUNDATIONS, THREATS & DEFENSE STRATEGIES

SkillWeed

CONTENTS

MODULE 1:
INTRODUCTION TO AI MODEL SECURITY



LESSON INTRODUCTION:

Welcome to the first module of our course on AI Model Security. As AI continues to drive decisions in finance, healthcare, law enforcement, and marketing, it becomes vital to ensure that these models are not only performant but secure, trustworthy, and ethical.

SECTION 1: WHAT IS AI MODEL SECURITY?

Definition:

AI Model Security is the set of practices and frameworks designed to protect AI/ML systems from threats that can compromise:

- The integrity of model predictions,
- The confidentiality of training and input data,
- The availability of the model itself,
- And the trustworthiness of its outputs.

Why It Matters:

- AI models operate on **sensitive data** (e.g., health records, financial transactions).
- They are vulnerable to **unique attacks** like adversarial examples and data poisoning.
- **Decisions made by AI models** are increasingly subject to **regulatory scrutiny**.

Key Insight:

Unlike traditional software, AI models learn from data. If the data is poisoned or manipulated, the model's predictions can be inaccurate, biased, or dangerous.

SECTION 2: THE AI/ML LIFECYCLE – WHERE SECURITY FITS

The AI lifecycle includes six core stages:

Stage	Security Concerns
1. Problem Definition	Misaligned objectives, ethical missteps
2. Data Collection	Data poisoning, copyright issues, biased datasets
3. Model Development	Model leakage, overfitting, malicious code injection
4. Model Evaluation	Improper testing, lack of fairness/robustness evaluation
5. Deployment	Adversarial attacks, model extraction, exposure of APIs
6. Monitoring & Updates	Data drift, concept drift, unauthorized model updates

SECTION 3: SECURITY GOALS FOR AI

In cybersecurity, we talk about the **CIA triad** — Confidentiality, Integrity, Availability — and AI expands this to include **Accountability** and **Explainability**:

- **Confidentiality**: Protect the training data, especially if it includes PII.
- **Integrity**: Ensure the model performs as expected and isn't altered or misled.
- **Availability**: Models must be resilient and accessible when needed.
- **Accountability**: Keep logs of decisions, access, changes.
- **Explainability**: Models should be understandable by humans to ensure trust and transparency.

Real-world Example: A healthcare diagnosis model should explain why it flagged a patient as high-risk, not just give a binary decision.

SECTION 4: THREAT LANDSCAPE FOR AI SYSTEMS

Let's look at the common categories of attacks specific to AI/ML:

Threat	Explanation
Adversarial Examples	Slightly modified inputs crafted to fool models (e.g., tricking a stop sign detector)
Data Poisoning	Corrupting the training set to produce inaccurate models
Model Inversion	Recovering sensitive training data from a model's outputs
Membership Inference	Determining if a specific individual's data was in the training set
Model Extraction	Copying a model by querying it repeatedly
Model Drift	The model becomes less effective as data patterns change over time

Discussion Prompt:

If someone can replicate your model just by using the API, what business risks could that present?

SECTION 5: KEY FRAMEWORKS AND GUIDELINES

Security of AI doesn't exist in isolation — it connects to privacy laws, ethical standards, and AI-specific governance. Key frameworks:

1. NIST AI Risk Management Framework (AI RMF):

- Covers four core functions: **Map, Measure, Manage, and Govern** risks in AI.
- Encourages a culture of accountability and transparency.

2. ISO/IEC 23894:2023:

- Guidance on managing risk for AI systems.
- Tied to ISO 27001 and cybersecurity management.

3. GDPR (General Data Protection Regulation):

- Applies to AI models processing data of EU citizens.
- Introduces requirements for explainability and right to opt-out of automated decisions.

4. OECD AI Principles:

- Promote AI systems that are **inclusive, sustainable, transparent, robust, and accountable**.

5. HIPAA (in Healthcare):

- Requires AI systems to ensure privacy and security of patient data.

SECTION 6: CASE STUDIES AND INDUSTRY EXAMPLES

Case 1: Adversarial Patch on a Stop Sign

Researchers added a sticker to a stop sign, and a self-driving car misclassified it as a yield sign — posing a deadly risk.

Case 2: Microsoft's AI Chatbot "Tay"

In less than 24 hours, Tay was manipulated through input poisoning to start producing racist and offensive tweets.

Case 3: Model Leakage in Healthcare

A predictive model accidentally exposed sensitive patient data through explainability outputs — violating HIPAA.

LAB EXERCISE: AI MODEL SECURITY THREAT MAPPING

Objective:

Map threats and controls across the AI/ML lifecycle using a template or digital whiteboard.

Instructions:

1. Download the lifecycle diagram template.
2. Identify at least **one security threat** and **one security control** at each stage.

Example Table Output:

Stage	Threat	Mitigation Strategy
Data Collection	Poisoned dataset	Data validation, provenance checks
Model Development	Backdoored code	Code review, static analysis tools
Deployment	Adversarial inputs	Input sanitization, adversarial training
Monitoring	Data drift	Continuous evaluation, alerting mechanisms

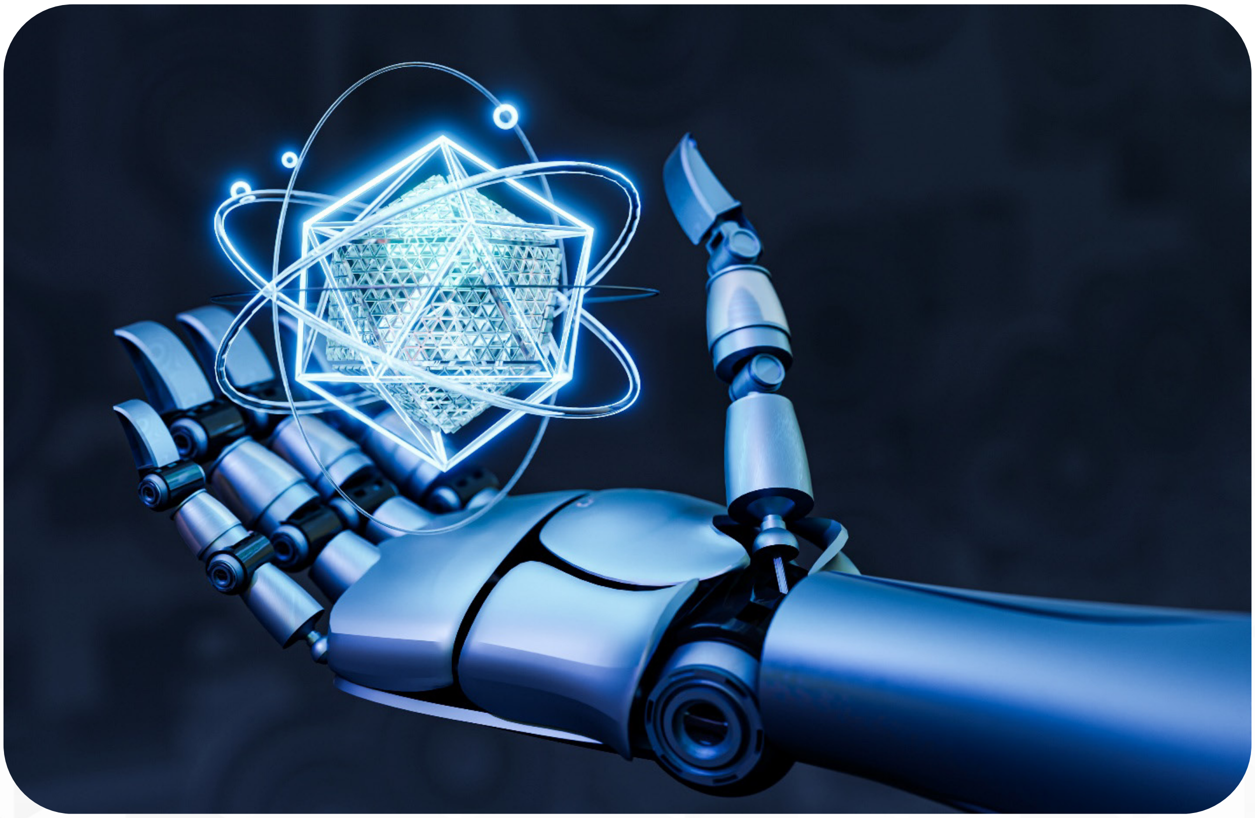
Quiz Questions (Self-Assessment)

1. What unique aspect of AI models makes them vulnerable to poisoning attacks?
2. Which framework defines the "Map, Measure, Manage, Govern" approach to AI risk?
3. How does GDPR affect automated decision-making in AI systems?
4. What is the difference between model inversion and membership inference?

CONCLUSION & TAKEAWAY

AI Model Security is not optional — it is foundational to trustworthy, legal, and effective AI systems. As we progress, we'll move from theory to hands-on security techniques and real-world defenses.

MODULE 2:
COMMON THREATS TO AI MODELS



LEARNING OBJECTIVES

By the end of this module, learners will be able to:

- Identify and describe common attack types on AI models
- Understand how and why these threats occur
- Recognize real-world examples of AI attacks
- Map threats to specific stages in the AI lifecycle
- Set the stage for applying defensive techniques in later modules

SECTION 1: INTRODUCTION TO AI THREATS

AI systems differ from traditional software in their **learning behavior**, **data dependencies**, and **lack of transparency**, which introduces new risks.

Key Vulnerabilities:

- Models are black-boxes
- Data is often scraped, labeled, or open source
- APIs are exposed for model inference
- Attackers can learn from model outputs

SECTION 2: DATA POISONING ATTACKS

Definition:
Attackers manipulate the training dataset to compromise the model's behavior.

Types:

- **Label Flipping:** Changing the labels of training examples (e.g., flipping 'safe' to 'unsafe').
- **Backdoor Injection:** Adding hidden triggers (e.g., if "sticker = yes," then classify as "authorized").

Example:

In facial recognition, inserting a specific pattern (e.g., sunglasses) can cause the model to misclassify a person as someone else.

Impact:

- Compromised integrity
- Biased or incorrect decision-making

Defense Preview:

Data validation, robust training, outlier detection

SECTION 3: ADVERSARIAL EXAMPLES

Definition:

Inputs that are intentionally perturbed in small, often imperceptible ways to mislead the AI model.

Example:

A stop sign with subtle noise gets classified as a speed limit sign.

Key Point:

These attacks exploit the geometry of high-dimensional space — even minor changes in pixel values can mislead a model.

Threat Vector	Target	Outcome
Image Pixels	Vision Model	Wrong classification
Text Embedding	NLP Model	Misinterpretation of meaning
Audio Signal	Speech AI	Incorrect transcription

Defense Preview:

Adversarial training, input sanitization, defensive distillation

SECTION 4: MODEL INVERSION ATTACKS

Definition:

An attacker uses model outputs to reconstruct or infer sensitive data from the training set.

Real-World Risk:

Recreating facial images used to train a public AI model or reconstructing private text data.

When It Happens:

- When APIs return rich confidence scores
- In white-box environments with internal access

Impact:

- Breach of confidentiality
- Violation of GDPR, HIPAA, etc.

Defense Preview:

Limit output details, add differential privacy, monitor usage patterns

SECTION 5: MEMBERSHIP INFERENCE ATTACKS (MIA)

Definition:

The attacker determines whether a specific data point was used in training.

Use Case:

In a hospital's AI model, an adversary could determine if a patient's record was included — a serious privacy breach.

Symptoms of Exposure:

- Overfitting (model behaves differently on known vs. unknown data)
- Overconfident predictions

Defense Preview:

Regularization, privacy-preserving ML, output obfuscation

SECTION 6: MODEL EXTRACTION ATTACKS

Definition:

An attacker recreates the model by repeatedly querying it and building their own approximation.

Why This Happens:

- Public APIs are often rate-unlimited
- Confidence scores provide too much info
- Attackers use shadow models

Business Risk:

Intellectual property theft — losing competitive edge

Famous Case:

Tramèr et al., 2016: Reconstructed commercial ML-as-a-service models using limited queries.

Defense Preview:

API rate limits, watermarking, response fuzzing

SECTION 7: MODEL DRIFT & DATA DRIFT

Definition:

Over time, real-world data shifts and the model becomes less accurate.

- **Data Drift:** Input features change
- **Concept Drift:** Target label relationships change

Example:

A fraud detection model trained on 2020 patterns may fail in 2025 due to new fraud techniques.

Security Impact:

- Increased false negatives
- Poor decision-making
- Exploitable by attackers

Defense Preview:

Ongoing monitoring, retraining pipelines, anomaly detection

SECTION 8: THREAT MAPPING TO LIFECYCLE

AI Lifecycle Stage	Threat Type	Description
Data Collection	Poisoning, Bias Injection	Malicious data inserted
Model Training	Backdoor Attacks, Overfitting	Triggers inserted during training
Model Evaluation	Over-optimization	Metrics gamed via overfitting
Deployment	Model Extraction, Drift	Exposure via API
Inference	Adversarial Inputs, MIA	Inputs crafted to manipulate results
Monitoring	Silent Failures, Drift	Lack of visibility post-deployment

HANDS-ON LAB: THREAT SIMULATION WORKSHOP

Goal:
Simulate or analyze 3 types of AI threats using real tools or diagrams.

Labs (Optional)

- Use Foolbox or Adversarial Robustness Toolbox
- Use pretrained model APIs (like Hugging Face) and simulate MIA using logs
- Alternatively, review and annotate published threat research (e.g., Google's adversarial ML examples)

Deliverable:

Submit a 1-page threat report describing the attack scenario, threat type, and one possible mitigation strategy.

QUIZ QUESTIONS:

1. What's the difference between a poisoning attack and an adversarial example?
2. What makes AI models vulnerable to model inversion?
3. How could an attacker perform a model extraction attack on a SaaS product?
4. Why does model overfitting make membership inference easier?

CONCLUSION & KEY TAKEAWAYS

- AI models introduce novel threats that go beyond traditional cybersecurity
- Understanding adversarial techniques is the first step to building robust AI
- Prevention requires secure design, careful data curation, and runtime monitoring
- These threats are not hypothetical — they are already happening across industries

MODULE 3: SECURE AI DEVELOPMENT LIFECYCLE (SAI-DLC)

LEARNING OBJECTIVES

By the end of this module, learners will be able to:

- Understand the concept of Secure AI Development Lifecycle (SAI-DLC)
- Identify security best practices for each AI development phase
- Apply privacy-preserving and explainability techniques to secure AI systems
- Build governance into the AI lifecycle for responsible AI adoption
- Design and apply security controls during development and deployment

SECTION 1: WHAT IS THE SECURE AI DEVELOPMENT LIFECYCLE (SAI-DLC)?

Definition:

The SAI-DLC is a structured approach that integrates security, privacy, fairness, and robustness throughout the AI system lifecycle — not just at the end.

Why It Matters:

- Embeds security early (shift-left)
- Reduces cost of vulnerabilities
- Aligns with DevSecOps and MLOps principles
- Ensures trustworthy and explainable AI outcomes

SECTION 2: SAI-DLC PHASES AND SECURITY CONSIDERATIONS

Phase	Key Security Concerns	Mitigation Techniques
1. Problem Definition	Ethical ambiguity, unaligned goals	Ethics review, stakeholder consultation
2. Data Collection	Poisoned or biased data, data leakage	Data validation, lineage tracking, anonymization
3. Model Development	Model theft, backdoors, bias	Secure coding, adversarial testing, fairness testing
4. Model Evaluation	Hidden vulnerabilities, overfitting	Robustness & explainability testing
5. Deployment	API abuse, model extraction, adversarial queries	Rate limiting, input filtering, watermarking
6. Monitoring & Feedback	Drift, unmonitored changes, silent failures	Logging, alerts, continual training loop

SECTION 3: SECURE DATA COLLECTION AND LABELING

Risks:

- Ingesting poisoned or fake data
- Use of copyrighted or private datasets
- Annotation errors or intentional labeling bias

Mitigations:

- **Data Provenance:** Verify the source of data
- **Anonymization & Pseudonymization:** Remove PII
- **Differential Privacy:** Add noise to protect individuals
- **Data Augmentation:** Improve data quality for resilience

Tool Examples:

- Google TensorFlow Privacy
- OpenMined's PySyft for federated learning

SECTION 4: SECURE MODEL DESIGN AND TRAINING

Risks:

- Use of unverified pre-trained models
- Insecure code in training pipelines
- Introduction of hidden triggers

Best Practices:

- **Secure Code Practices:** Use code reviews, linters, and secrets scanning
- **Use Trusted Libraries:** Avoid obscure or unverifiable frameworks
- **Adversarial Training:** Train on adversarial examples to boost resilience
- **Fairness Testing:** Audit for bias using tools like IBM AI Fairness 360

Example:
Training a model on salary prediction that removes gender or race bias via preprocessing.

SECTION 5: ROBUST MODEL EVALUATION

Risks:

- Overfitting to test data
- False sense of robustness
- Hidden vulnerabilities (e.g., in edge cases)

Controls:

- **Cross-validation with noise-injected datasets**
- **Stress Testing & Perturbation Tests**
- **Explainability Checks:** Use SHAP, LIME, or GradCAM

Explainability in Evaluation:

- Ensures models are not making decisions based on unintended correlations (e.g., a radiology AI model using surgical markers instead of pathology).

SECTION 6: SECURE MODEL DEPLOYMENT

Risks:

- API scraping leading to model theft
- Model tampering or unauthorized updates
- Injection attacks via input fields

Secure Deployment Tactics:

- **Model Watermarking:** Embed unique markers to detect stolen models
- **Rate Limiting and Throttling:** Prevent brute force and extraction
- **Input Sanitization:** Cleanse input data at API endpoints
- **Containerization & Secrets Management:** Use Docker/Kubernetes with secure keys

Recommended Tools:

- AWS SageMaker with IAM roles
- Azure ML with private endpoints

SECTION 7: CONTINUOUS MONITORING AND GOVERNANCE

Why It's Critical:

- AI performance and behavior drift over time
- Undetected anomalies can lead to systemic failure

Monitoring Strategy:

- Set performance thresholds and alerts
- Use dashboards to visualize drift and bias
- Enable rollback mechanisms for misbehaving models

Governance Layer:

- Assign clear ownership for model decisions
- Use model version control (e.g., MLflow, DVC)
- Ensure model logs are auditable and compliant

SECTION 8: HUMAN-IN-THE-LOOP (HITL) FOR RISK MITIGATION

- **What It Is:** Keeping a human decision-maker in the loop during model training, decision-making, or re-training
- **Benefits:** Better oversight, ethical control, and transparency
- **When to Use:** High-risk domains (e.g., healthcare, justice, lending)

LAB EXERCISE: BUILDING A SAI-DLC CHECKLIST (OPTIONAL)

Objective:

Create a secure development checklist for your AI project.

Instructions:

1. Pick a use case (e.g., fraud detection, medical diagnosis, facial recognition)
2. Map out the development lifecycle
3. For each stage, list 1–2 security best practices
4. Share your checklist with the class or submit it for feedback

Example Output:

Lifecycle Stage	Security Practice
Data Collection	Remove PII, validate source
Model Training	Use adversarial training, code review
Deployment	Enable API monitoring, add watermark
Monitoring	Track data drift, auto-flag anomalies

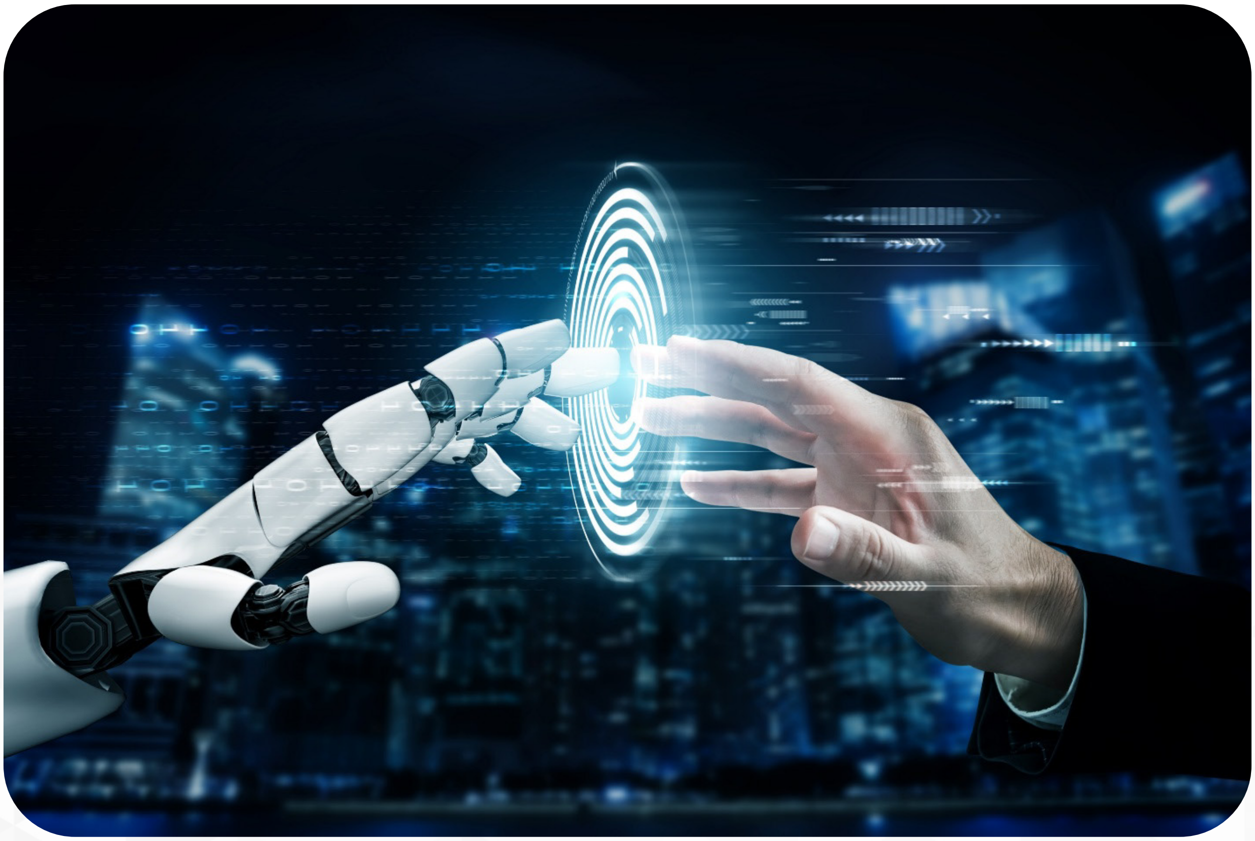
Quiz Questions:

1. What is the difference between adversarial training and differential privacy?
2. Why is explainability important in model evaluation?
3. What security controls can prevent model extraction?
4. List two reasons why securing the data pipeline is critical to AI model security.

CONCLUSION AND TAKEAWAYS

- Security must be woven into each phase of AI development
- Adopting SAI-DLC reduces risks, costs, and regulatory exposure
- Tools like adversarial training, explainability, privacy-preserving ML, and monitoring form a holistic defense
- The future of AI security lies in proactive design, not reactive patches

MODULE 4:
AI GOVERNANCE AND COMPLIANCE



LEARNING OBJECTIVES

By the end of this module, learners will be able to:

- Define AI governance and its importance in managing AI risks
- Understand core regulatory and ethical principles governing AI use
- Apply compliance standards such as GDPR, HIPAA, and NIST AI RMF
- Establish AI oversight mechanisms, including risk registers and model audits
- Evaluate trade-offs between explainability, fairness, and performance

SECTION 1: WHAT IS AI GOVERNANCE?

Definition:

AI Governance is the framework of policies, processes, roles, and tools that ensure AI systems are **ethical, accountable, secure, and compliant** with laws and organizational standards.

Purpose:

- Reduce risk exposure
- Ensure AI aligns with human values and societal goals
- Meet legal, regulatory, and industry-specific obligations

Three Pillars of AI Governance:

1. **Transparency** – Knowing how and why AI makes decisions
2. **Accountability** – Assigning ownership for AI actions
3. **Trustworthiness** – Ensuring AI is safe, secure, and fair

SECTION 2: AI COMPLIANCE FRAMEWORKS AND STANDARDS

Framework / Regulation	Region / Focus	Key Focus Areas
NIST AI RMF (2023)	U.S., Government & Industry	Risk-based approach to AI design & deployment
ISO/IEC 23894:2023	Global (ISO)	Risk mgmt framework for AI systems
GDPR (Article 22)	EU	Automated decision-making, data subject rights
AI Act (EU)	EU	Risk-classification of AI systems
OECD AI Principles	38 Countries incl. U.S. and EU	Inclusive, sustainable, human-centered AI
HIPAA (USA)	Healthcare	PHI security and privacy
CPRA / CCPA (California)	U.S., California	Personal data rights and transparency

SECTION 3: KEY COMPLIANCE CONCEPTS

1. Lawfulness of Processing

- Must have legal basis to collect and process data
- Consent or legitimate interest is required (GDPR, HIPAA)

2. Purpose Limitation

- Data should only be used for the purpose it was collected for

3. Fairness and Non-discrimination

- Avoid bias in AI systems
- Perform fairness testing (race, gender, socio-economic factors)

4. Transparency and Explainability

- Users and stakeholders must understand how AI decisions are made
- Required in EU's GDPR and AI Act

5. Human Oversight

- High-risk decisions (e.g., health, loans, hiring) must allow human intervention

6. Data Minimization

- Only necessary data should be collected and retained

SECTION 4: GOVERNANCE STRUCTURES & ROLES

Role	Responsibility
Chief AI Ethics Officer	Oversees ethics, bias audits, and fairness policies
AI Model Risk Committee	Approves deployment of sensitive AI systems
Data Protection Officer (DPO)	Ensures compliance with privacy laws
ML Engineer / Scientist	Implements technical controls for compliance
AI Auditor	Conducts explainability and impact assessments

Tip: Organizations should document AI system use in **AI Risk Registers** and **Model Cards** that include purpose, dataset, risk rating, and known limitations.

SECTION 5: EXPLAINABILITY VS. PERFORMANCE TRADE-OFF

Challenge:

Complex models like deep neural networks provide higher accuracy, but lower interpretability. Regulations like GDPR and the EU AI Act require explanations for automated decisions.

Model Type	Accuracy	Explainability
Decision Trees	Medium	High
Random Forests	High	Medium
Deep Neural Networks	Very High	Low
Rule-Based Systems	Medium	Very High

Solution Strategies:

- Post-hoc explainability (e.g., SHAP, LIME)
- Use of surrogate models
- Model documentation and disclaimers

SECTION 6: IMPLEMENTING AI GOVERNANCE POLICIES

Organizations should codify AI governance via policies:

1. **AI Use Policy** – Defines acceptable use of AI tools
2. **Model Lifecycle Policy** – Documents roles, review, and change procedures
3. **Bias Mitigation Policy** – Describes procedures for testing and remediating bias
4. **Privacy Impact Assessments (PIA)** – Mandatory for high-risk use cases (e.g., health, finance)

SECTION 7: AUDITING AND COMPLIANCE MECHANISMS

Model Auditing Checklist:

- Is the model explainable and documented?
- Are data sources ethical and compliant?
- Has the model been tested for bias and drift?
- Is there a rollback plan if the model fails?
- Who is accountable for model decisions?

Toolkits for Governance:

- **IBM AI Fairness 360**
- **Google What-If Tool**
- **Microsoft Responsible AI Dashboard**
- **AI Explainability 360 Toolkit**

LAB EXERCISE: AI COMPLIANCE AUDIT REPORT (OPTIONAL)

Scenario:

You are the AI Compliance Officer in a healthcare company using an AI model for diagnostic predictions.

Task:

1. Complete a mini-compliance audit based on a checklist:
    - Is the data anonymized?
    - Are patients informed of AI use?
    - Is there a human-in-the-loop mechanism?
    - Have you tested for fairness and drift?
2. Submit a short 1-page audit summary including:
    - Areas of compliance
    - Areas needing mitigation
    - Recommended next steps

QUIZ QUESTIONS:

1. What key AI principle is embedded in GDPR's Article 22?
2. Name one tool that supports explainability in AI models.
3. Why is model documentation (e.g., model cards) important for governance?
4. What's the role of an AI Ethics Officer in governance?

CONCLUSION & KEY TAKEAWAYS

- AI governance ensures responsible, fair, and secure use of AI
- Compliance with global laws (GDPR, HIPAA, AI Act) is a legal and ethical obligation
- Governance requires both **policy structure** and **technical enforcement**
- Explainability, bias testing, and human oversight are foundational
- AI security is not complete without governance

APPENDIX 1:
FINAL PROJECT: AI SECURITY RISK ASSESSMENT AND SECURE LIFECYCLE DESIGN (OPTIONAL)

PROJECT OBJECTIVE

The goal of this final project is to apply the knowledge gained throughout the course to evaluate the security posture of a real or hypothetical AI system and develop a secure development lifecycle roadmap (SAI-DLC) for that system.

PROJECT SCENARIO OPTIONS (PICK ONE OR PROPOSE YOUR OWN):

1. **Healthcare** – AI model for early cancer detection using patient imaging data.
2. **Finance** – AI model for loan approval and fraud detection.
3. **Retail** – AI recommendation system for e-commerce personalization.
4. **Education** – AI grading tool for automated student assessments.
5. **Proposed System** – Choose a real system you're working on or researching.

PROJECT DELIVERABLES

1. **System Overview (1–2 pages)**
    - Describe the AI system, its purpose, users, and environment.
    - Identify the type of model used (e.g., CNN, NLP, ensemble).
2. **Threat Landscape Analysis**
    - Identify at least 5 security threats specific to this system.
    - Categorize threats (e.g., poisoning, inversion, adversarial).

3. **Secure AI Development Lifecycle (SAI-DLC) Roadmap**
    - Define security actions at each stage:
        - Problem definition
        - Data collection
        - Model training
        - Evaluation
        - Deployment
        - Monitoring
4. **Compliance & Governance Plan**
    - Identify applicable regulations (GDPR, HIPAA, NIST AI RMF, etc.).
    - Describe how explainability, fairness, and accountability will be enforced.

APPENDIX 2: SAMPLE HIGH-QUALITY STUDENT PROJECT RESPONSE

PROJECT TITLE:

Securing AI-Powered Diagnostic Imaging in a Telehealth Platform

1. SYSTEM OVERVIEW

Use Case:

An AI system embedded in a telehealth platform is used to analyze X-ray and CT scan images to detect early signs of lung cancer. It provides diagnostic recommendations to radiologists.

Model Type:

Convolutional Neural Network (CNN), trained on labeled imaging data sourced from multiple hospitals and open datasets.

Stakeholders:

Radiologists, Patients, Medical Researchers, Compliance Officers, Telehealth Engineers

2. THREAT LANDSCAPE ANALYSIS

Threat	Description
Data Poisoning	Malicious actors may upload altered images that train the model to ignore tumors
Adversarial Examples	Input images are subtly modified to trick the model into missing a diagnosis
Membership Inference	Attackers determine if a patient's scan was in the training set (privacy breach)
Model Inversion	Reconstructing patient imaging data from model responses
Model Drift	Over time, model accuracy degrades as new imaging tech or diseases evolve

3. SECURE AI DEVELOPMENT LIFECYCLE (SAI-DLC)

Stage	Security Actions
Problem Definition	Ensure ethical guidelines align with healthcare data protection and equity
Data Collection	Use HIPAA-compliant sources; anonymize data; implement lineage tracking
Model Training	Adversarial training; static code analysis; restrict access to training scripts
Evaluation	Run robustness and fairness testing across demographics; explainability with LIME
Deployment	Deploy behind secure APIs; watermark model to detect theft; restrict access via IAM
Monitoring	Implement drift detection; log anomalies; retrain quarterly with updated datasets

4. COMPLIANCE & GOVERNANCE STRATEGY

- **HIPAA:** All patient imaging data is de-identified and access-controlled.
- **GDPR Article 22:** Patients have the right to request human review of AI-assisted decisions.
- **NIST AI RMF:** Risks are mapped and mitigated using the "Map-Measure-Manage-Govern" cycle.
- **Governance:** AI Ethics Officer oversees deployments, and all model updates are version-controlled and auditable.

FINAL COMMENTS

This AI system requires continuous oversight to ensure model accuracy, ethical use, and patient trust. By embedding security into the full development lifecycle, we mitigate risks that could result in harm to patients and reputational damage to the platform.

SkillWeed