Day 11

Generative AI
Security

Trustworthiness Characteristics – NIST AI RMF:

These are the qualities that make an AI system "trustworthy." They ensure that the model behaves safely, ethically, and reliably across its lifecycle.

Safety: Safety means making sure the AI does not cause harm — intentionally or unintentionally.

What it includes:

- Preventing harmful outputs
  Example: A chatbot must not generate self-harm instructions or violent content.
- Failure-mode analysis
  Studying how the model fails (e.g., hallucinations, toxic content).
- Safe deployment
  Using human review for high-risk decisions (healthcare, finance).

Example: A medical diagnosis model is tested to make sure it never suggests unsafe medicines.

Security: Security means protecting the AI system from attacks.

What it includes:

- Poisoning attacks
  Attackers modify training data so the model learns incorrect patterns.
- Model extraction
  Attackers reverse-engineer the model to steal it.
- Evasion attacks
  Slightly modify input to fool the model (e.g., adversarial images).
- LLM security risks
  Jailbreak prompts, prompt injection, data leakage.

Example: Adding adversarial training to prevent someone from tricking a face-recognition model with a sticker

Privacy: Protecting personal data used in training or generated by the model.

What it includes

- Data minimization
  Collect only the data needed.
- Anonymization / de-identification
  Removing identifiable fields.
- Secure training & storage
  Encrypt datasets, avoid PII leakage.

- Privacy-preserving ML
  Differential privacy, federated learning.

Example: Training a fraud detection model on anonymized transaction data so no user identity is revealed.

Fairness: Fairness means the AI should not discriminate against any group.

What it includes

- Bias detection methods
  Measuring if certain groups get worse predictions.
- Fairness metrics
  - Demographic parity
  - Equal opportunity
  - Equalized odds
- Bias mitigation
  Reweighting data, debiasing algorithms.

Example: Checking if a loan approval model rejects more applications from a certain gender or community.

Transparency: Transparency means making the AI system understandable.

What it includes

- Documentation
  Data sheets, model cards, decision logs.
- Explainability
  Methods like SHAP, LIME to explain predictions.
- Clear user communication
  The user must know when AI is being used.

Example: A model card showing what data was used, its limitations, and safe-use guidelines.

Accountability: Accountability means you can trace decisions and identify who is responsible.

What it includes

- Auditability
  Logs of inputs, outputs, and model changes.
- Traceability
  Track which dataset, which version of the model, which parameters produced the output.
- Clear role responsibilities
  Who approves models, who updates them, who monitors them.

Example: Keeping audit logs so that if an AI system makes a wrong medical decision, you can trace what happened.

Reliability: Reliability means the AI works consistently across different conditions and over time.

What it includes

- Stress testing
  Test the model with edge cases.
- Performance consistency
  Check accuracy in different environments, languages, user behaviors.
- Robustness metrics
  Accuracy drop under noise, adversarial robustness, drift performance.

Example: A speech recognition model must work in quiet rooms and noisy streets.

Basically,

- Safety → Don’t harm people, avoid dangerous outputs.
- Security → Protect model from attacks.
- Privacy → Protect personal data and avoid leakage.
- Fairness → No discrimination, unbiased results.
- Transparency → Clear documentation, explainability.
- Accountability → Clear responsibilities, audit logs, traceability.
- Reliability → Stable and consistent performance across all conditions.

AI Cybersecurity Profiles (AI RMF):

What are cybersecurity profiles?

A cybersecurity profile is a structured document that describes:

- the type of AI system,
- the key risks it faces, and
- the controls required to protect it.

It acts like a template or blueprint for securing a specific category of AI systems.

Simple example:

A "Foundation LLM Profile" may list:

- Risks → jailbreaks, data leakage, hallucination
- Required controls → content filters, red teaming, access control

Why do we use profiles?

Profiles help an organization:

- apply consistent security standards
- improve faster risk assessments
- identify missing controls
- support audits and compliance
- guide security teams and ML teams with a common reference

How to build an AI profile for your organization

Creating a profile typically follows these steps:

Step 1: Identify the AI system type

Step 2: Define the system purpose and use case

Step 3: Identify risks

Step 4: Map threats to controls

Step 5: Define security requirements

Step 6: Document lifecycle processes

Step 7: Review and approve

Example Profiles:

A. Foundation Model Profile
B. RAG (Retrieval-Augmented Generation) Profile
C. Vision Model Profile

Mapping threats & controls to profiles: This is a core RMF activity.

Example:

Threat	**Profile**	**Control**
Prompt injection	LLM / RAG	Input sanitization, layered guardrails
Model extraction	Foundation	Rate limiting, watermarking
Retrieval poisoning	RAG	Content validation pipeline
Bias	All	Fairness metrics, bias testing
Privacy leakage	All	Differential privacy, redaction

Using Profiles for Audits & Compliance Reviews:

Profiles act as a checklist during audits.

How they help audits

- Show what risks were identified
- Show controls implemented
- Provide justification for missing controls
- Make governance more transparent
- Demonstrate compliance with NIST, ISO 42001, EU AI Act

Basically,

- Cybersecurity profiles → Templates that describe risks + controls for each type of AI system.
- Profiles make security consistent across the organization.
- Examples → Foundation Model Profile, RAG Profile.
- Threat–control mapping helps choose protections for each risk.
- Profiles help in audits because they act like a checklist for compliance.

--The End--