Day 10

NIST AI Risk Management Framework (AI RMF):

1. **Core Function: MEASURE**

MEASURE = Test, evaluate, stress, and quantify the risks of your AI system. If MAP tells you "what can go wrong,". MEASURE tells you "how bad it is, how likely it is, and where your model fails."

This function is all about evaluations, scoring, testing, and measurement.

Risk Assessments: checking what risks exist and how serious they are.

Activities include:

- Identify harms (bias, hallucination, misuse, attacks)
- Score likelihood (low/medium/high)
- Score impact (minor/medium/critical)
- Combine them into a risk rating

Example:

- Likelihood: Prompt injection = High
- Impact: Could expose secrets = Critical
  → Risk = High

Why it matters:

- Helps prioritize what needs fixing
- Supports decision-making and approvals

Quantitative / Qualitative Evaluations: Two ways to measure AI risks:

a. Qualitative: human judgment is needed, no numeric score exists.
b. Quantitative: use metrics, scores, formulas.

Red Teaming and Adversarial Testing:

Red teaming = trying to break the model intentionally. Goal: simulate real attacks or misuse. Types of red teaming: Prompt-based attacks, Adversarial inputs, Misuse testing.

Security Testing Methods for LLMs:

LLMs need specialized security tests: prompt injection testing, model extraction testing, data leakage testing, Jailbreak testing, Abuse testing.

Bias Assessments: Bias assessment checks whether the model treats groups fairly.

Bias metrics may include:

- Demographic parity
- Equal opportunity score
- False Positive Rate by group
- False Negative Rate by group
- Toxicity differences across populations

Bias can be measured at:

- Data level
- Model prediction level
- Output text level

Safety and Content Evaluation: Safety evaluation ensures the model does not produce harmful or unsafe content.

Content categories tested:

- Hate speech
- Violence
- Self-harm
- Extremism
- Sexually explicit content
- Misinformation
- Disallowed medical/legal advice

Methods include:

- Automated safety classifiers
- Human evaluators
- Real-time monitoring

Evaluation Metrics for Robustness: Robustness = model's ability to work reliably under stress, attacks, and variations. Robustness metrics include:

- Tolerance to adversarial examples
- Stability across data shifts
- Drift detection metrics
- Consistency under repeated queries
- Jailbreak resistance score
- Success rate of adversarial attacks

Why robustness matters: protects against attacks, prevents unexpected failures, improves reliability in real-world conditions.

Basically, MEASURE = test the AI system to find problems, bias, attacks, and weaknesses.

You learn to:

a. Do risk assessments
b. Perform qualitative and quantitative evaluations
c. Conduct red teaming
d. Apply security testing for LLMs
e. Check bias
f. Test safety and harmful content
g. Measure robustness

Basically: MEASURE = "Stress test the AI system to see where it breaks and how safe it is."

2. Core Function: MANAGE

MANAGE = Take action to reduce AI risks and keep the system safe over time. If MEASURE finds the problems, MANAGE is about fixing them, monitoring them, and improving continuously. This function ensures that AI systems remain safe, secure, compliant, and trustworthy throughout their lifecycle.

Risk Prioritization: Not every risk is equally important.

Prioritization means identifying: high impact, high likelihood, require immediate fixes, can be accepted or postponed.

Common method: red, yellow, green.

Factors used to prioritize:

- Severity of harm
- Potential misuse
- Regulatory impact (EU AI Act, GDPR violations etc.)
- Business impact
- Reputational harm

Applying Controls: Controls = protections or safeguards added to reduce risks.

Examples of controls:

- Security controls
- Privacy controls
- Ethical controls
- Operational controls

Monitoring Risk Over Time: AI systems change over time due to:

- New data
- New user inputs
- New threats
- Model drift
- Changing business/legal requirements

So, risks must be monitored continuously, not once.

Monitoring includes:

Real-time monitoring

- Unsafe outputs
- Prompt injection attempts
- Bias spikes
- Sudden performance drop
- Data drift
- Model behavior anomalies

Regular audits

- Weekly or monthly reviews
- Check logs
- Validate performance
- Security audits

Updating Risk Assessments:

As models evolve, so do the risks. Risk assessments must be updated when:

- The model is retrained
- New features are added
- New datasets are used
- A new vulnerability is discovered
- Attack patterns evolve
- Regulations change
- A security incident occurs

Managing Vendor/Supplier Risks:

Many organizations use:

- Third-party LLMs (OpenAI, Anthropic, Google)
- External datasets
- Third-party APIs
- Vector databases
- AI platforms and SaaS tools

Vendor risks include:

- Model vulnerabilities
- Data privacy issues
- Unknown training data
- Compliance failures
- Supply chain attacks
- Poor documentation

Vendor risk management activities:

- Security review of vendors
- Contractual safeguards
- SLA requirements
- Red teaming external APIs
- Monitoring third-party updates
- Ensuring vendor aligns with your AI governance policies

Change Management in AI Systems

AI systems change continuously due to:

- retraining
- fine-tuning
- new tools
- new embeddings
- new prompts
- new infrastructure
- new use cases

Change management ensures safe and controlled updates.

Basically, MANAGE = Fix the risks and keep AI safe over time.

You learn to:

- Prioritize risks
- Apply controls to reduce them
- Monitor behavior continuously
- Update risk assessments as system changes
- Manage vendor risks
- Use change management to safely update models
- MANAGE = “Take action, fix issues, and make sure AI stays safe forever.”

--The End--