Microsoft Security

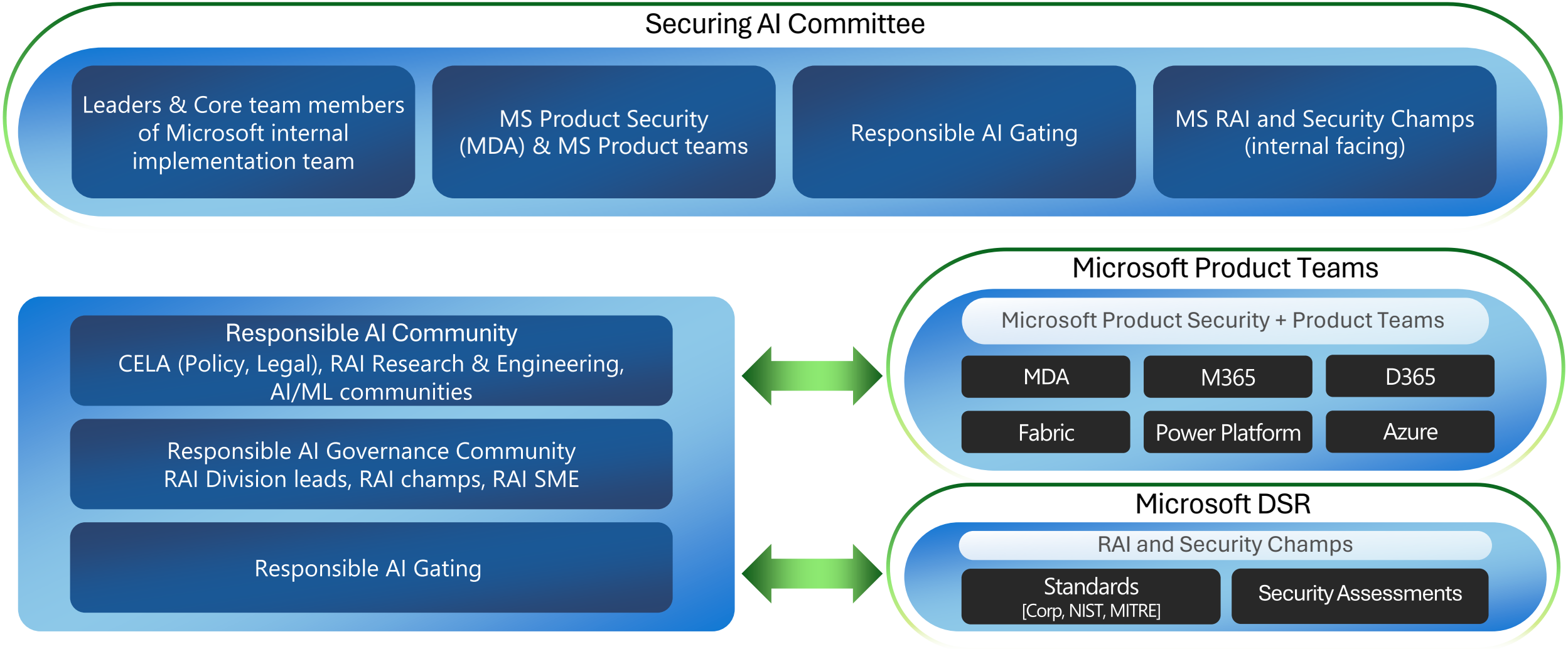
NIST Secure Use of LLMs and Generative AI Systems

Vivek Vinod Sharma
Senior Security Architect

Agenda

- Programmatic Overview
- Managing the development of GenAI
- Managing the deployment of GenAI
- Takeaways and Actions

Programmatic Overview for Securing AI at Microsoft



Securing Generative AI Program Strategy

Security assessment for first-party AI systems

Reduce risk of data exposure, insider threat, misuse before they are enabled for the tenant by maintaining security and compliance

Consumer risk reduction

Reduce risk of improper use of LLMs and LLM tools by providing employee guidance and education

Improving how Microsoft develops new AI systems

Enforce Microsoft Security AI Standards

Guiding principles – SD3 (Secure By Design, Secure By Default, Secure By Deployment)
Enable product teams to build products securely by providing standards and guidance via SDL and RAI

Securing Azure Tenant AI/ML workspaces

Reduce the risk of data exposure, misuse, abuse, insider threat of 80K subscriptions

Protect Sensitive data handling within AI apps

Reduce the risk of data exposure when using third party apps and services

Training and Awareness

Amplify AI security awareness across the company by providing trainings.

Reinforcing existing products, processes and systems

Our process to manage developing Generative AI

Submission

Centralized Intake and Triage
- Processed at the organization level
- Reviewed by centralized team
- Appropriate teams involved (internal vs. external)

Assessment

Completion of assessments
- RAI Impact Assessment
- Security Assessment
- SDL Requirements

Reassessment
Triggered when:
- Compliance failure
- Vulnerability discoveries
- Changes to policy
- Bi-annual re-attestation

Review

Review of assessments
- RAI and Security assessments: go/no-go
- RAI Gating team provides approval for different release phases

Our process to manage deploying Generative AI

Intake

Centralized Intake and Triage
Processed at the company-wide level
Open to all feature / product teams
Includes representatives from Engineering, Compliance, Security, Privacy, Legal

Assess

Completion of assessments
Criteria aligned to Enterprise Risk Management framework
Risks and gaps identified and discussed

Mitigate

Completion of Risk Mitigation Activities
Standards and policy enforcement
Approval of exceptions from senior leadership

Deploy

Deployment
Gated deployment triggered by tenant administrator
Publish or update employee-facing guidance / training

Key takeaways and actions

1. **Data Protection:** Apply labels and blocks to sensitive information not authorized for GenAI systems. Limit access appropriately
2. **Prompt Injection:** Inspect prompts and look for signs of abuse – including in code
3. **Data Poisoning:** Ensure validation of skills, functions, and plugins to prevent system abuse. Ensure appropriate configurations
4. **Audit & Logging:** Ensure all user activities, application activities are audited and logged
5. **Threat Monitoring & Response:** Ensure threat modeling, monitoring, and response are present. Conduct adversarial testing.