Learn about the New NIST SP 800-53 Control Overlays for Securing AI Systems Project

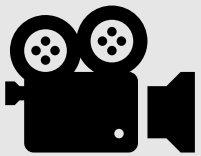
Virtual event will start at 1:00 PM ET

NIST | NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Virtual Event Etiquette & Reminders

Due to the number of attendees, all **audience** microphones and cameras are muted

Please enter questions and comments for presenters in the Q&A.



This virtual event is being recorded and will be posted in about 10 business days.

For issues, questions or comments, please use the **Zoom Q&A feature** or email: overlays-securing-ai@list.nist.gov

Virtual Event FAQ

Will this event be recorded?

Yes. This virtual event will be recorded and posted in about 10 business days on the event site.

When will slides be posted?

Slides will be posted immediately after the event.

Can I get CEUs for attending?

You may elect to self-certify you attended this event and use your registration confirmation email. NIST does not provide any forms or attestations.

Where is the event site?

https://csrc.nist.gov/Events/2025/ai-overlay-webinar

Learn about the New NIST SP 800-53 Control Overlays for Securing AI Systems Project

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Victoria Pillitteri
victoria.pillitteri@nist.gov

Apostol Vassilev
apostol.vassilev@nist.gov

Agenda

01 The Need: Cybersecurity Guidelines for Artificial Intelligence Systems

02 NIST Overview: Roles in Cybersecurity, Privacy, and Artificial Intelligence

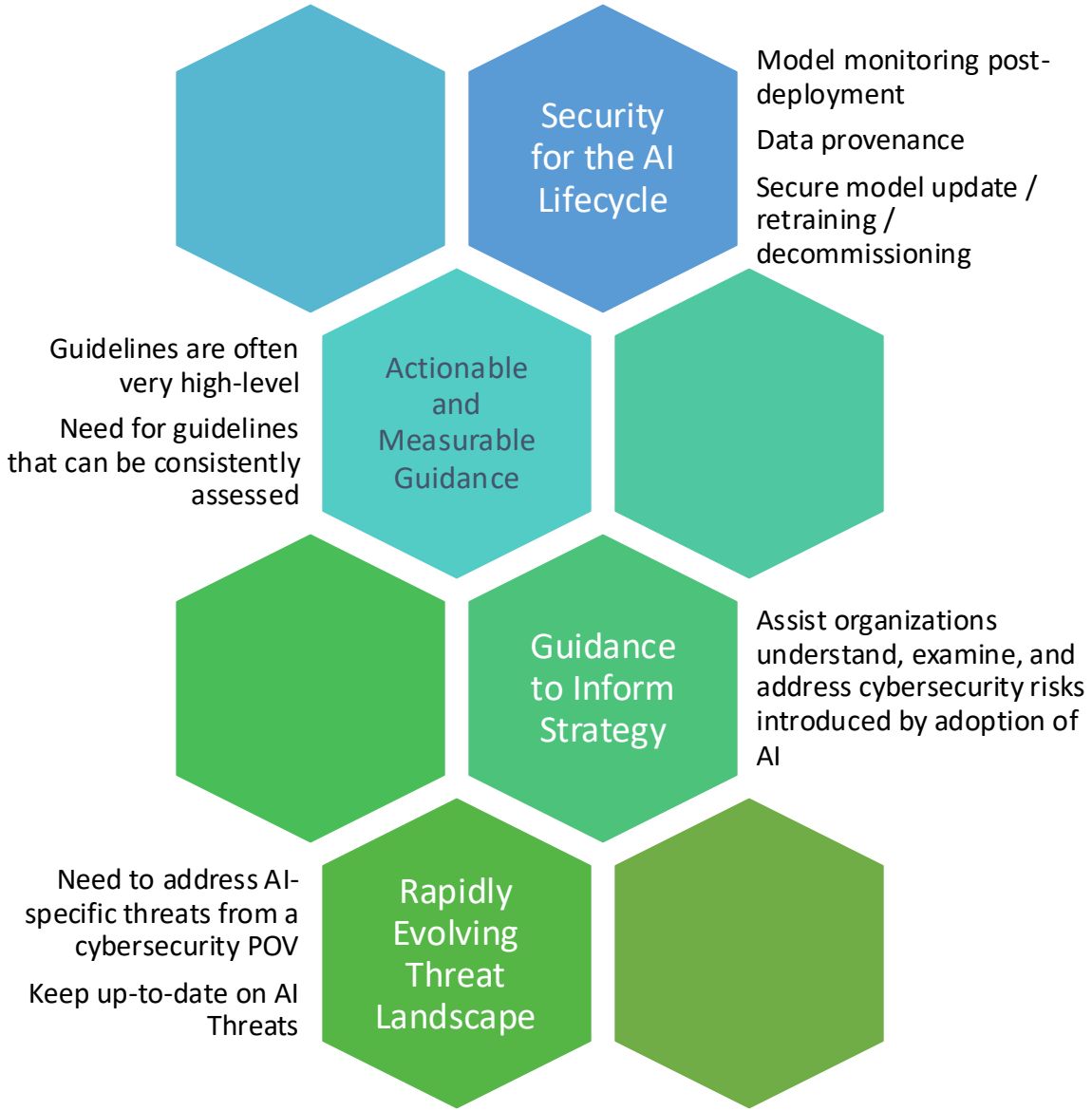
Featuring: Martin Stanley (AI RMF), Kat Megas (CyberAI Profile), and Maia Hamin (CAISI)

03 Addressing a need: Control Overlays for Securing AI Systems

04 Next Steps and How to Get Engaged
Q&A

The Need: Cybersecurity Guidelines for AI Systems

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE



NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Driving Innovation

From developing precise atomic clocks to creating encryption standards to supporting manufacturing, NIST plays a critical role in **advancing technology and US competitiveness** through research in measurement science and standards development

The Newton apple tree on the NIST Gaithersburg campus || J. Stoughton/NIST(2018)

Role in Cybersecurity & Privacy

- Conduct **research, development and outreach** to provide **standards, guidelines, mechanisms, tools, metrics and practices** to protect US information and systems.
- **Implement practical cybersecurity and privacy** through outreach and effective application of standards and best practices to support U.S. adoption of cybersecurity capabilities.

Role in AI

- **Promote innovation** and **cultivate trust** in the **design, development, use** and **governance** of **AI systems**
- **Facilitate testing** and **collaborative research** related to **harnessing and securing** the potential of commercial **AI systems**

AI Risk Management Framework

The Need: Trustworthy AI Cybersecurity Capabilities

Current Work: Planning Phase for an AI RMF Profile for Cybersecurity

Next steps: Draft a Concept Note to begin engagement with the community

Learn More and Get Engaged!

https://airc.nist.gov

Cyber AI Profile

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Organizations vary on whether and how they are using AI in these three areas. Some organizations may not yet be using AI. Regardless of where they are on their AI journey, organizations need risk management approaches that support the realities of advancements in AI use.

Securing AI ***in*** the enterprise (***Secure***)

Challenges in Protecting Enterprise AI

- Data Governance, Security, and Privacy
- Subject to Adversarial Attacks
- Unauthorized Access and Use
- Identifying, Detecting, & Responding to AI Component Vulnerabilities and Adverse Events
- Supply Chain Security
- Model Drift & Unexpected or Inaccurate Results

Using AI to protect the enterprise (***Defend***)

Opportunities for AI in Cyber Defense

- Advanced Threat Detection
- Advanced Threat Analysis
- Automated Incident Response
- Proactive Risk Management
- Security Governance & Policy

Protecting the enterprise ***from*** AI (***Thwart***)

New or Augmented Threats

- Automated & Adaptive Malware
- Targeted Phishing & Social Engineering
- AI-Driven Cyber Espionnage
- Evasion Techniques
- Supply Chain Attacks
- AI-Powered Zero-Day Exploits
- AI-Powered Attack Automation

For more information:
https://www.nccoe.nist.gov/projects/cyber-ai-profile

Center for AI Standards and Innovation

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

On June 3, 2025, Commerce Secretary Lutnick re-formed CAISI to:

1. **Lead evaluations and assessments of U.S. and adversary AI systems,** including adoption of foreign AI models and potential security vulnerabilities or malign foreign influence.
2. **Establish voluntary agreements with AI developers** to evaluate AI capabilities with national security implications.
3. **Develop best practices for improving AI security**, in collaboration with other NIST organizations.
4. **Coordinate with other federal agencies** to develop and conduct evaluations.
5. **Guard against burdensome and unnecessary regulation** by foreign governments and ensure dominance of U.S. AI standards.

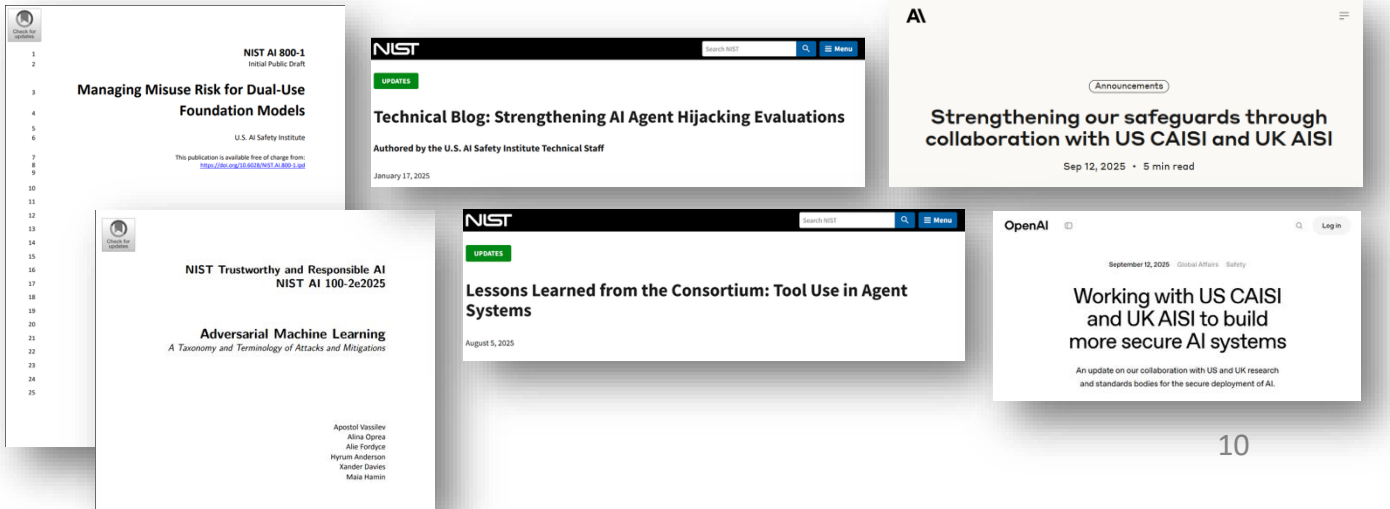
CAISI has:

Completed 10 voluntary pre-deployment national security tests of OpenAI's o1, o3-mini, o3, o4-mini, GPT-4.5 and GPT-5; Anthropic's Sonnet 3.5, 3.7, and 4 and Opus 4; xAI's Grok 3; plus post-deployment evaluations of 20+ U.S. and PRC models.

Signed research and testing agreements with major AI developers, third-party evaluators, and universities, plus general research agreements with 290+ private AI orgs in the AISI Consortium.

Directly collaborated on national security tests with 6+ federal agencies

Developed and collaborated on best practices and voluntary guidelines for evaluations and AI security



https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute-consortium-aisic

Control Overlays for Securing AI Systems

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

The controls to manage cybersecurity risks to AI systems will ***largely be similar*** to those required for any type of software.
Many organizations are ***familiar with the SP 800-53 controls*** and may already be implementing them.
The SP 800-53 controls offer ***flexibility*** to meet the ***unique security considerations for AI systems***.

COMMON **TECHNICAL FOUNDATION** FOR CYBERSECURITY OUTCOMES

IMPLEMENTATION-FOCUSED FOR DIFFERENT **AI USE CASES**

CAN LEVERAGE EXISTING **ASSESSMENT** GUIDELINES (SP 800-53A)

PROVIDES LINKS TO OTHER KEY CYBER & AI **NIST PUBS**

https://csrc.nist.gov/Projects/cosais overlays-securing-AI@list.nist.gov

What are Control Overlays?

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

What

- Additional customization options
- Can be a fully specified set of controls, or a subset to address a specific need

Why

- Opportunity to add, modify, or eliminate controls from a baseline
- Specify a set of controls for a specific need/scope (technology, environment of operation, type of system or mission, industry sector, etc.)
- Identify parameter values

Who

- Any stakeholder or community of interest!

NIST SP 800-53B provides additional guidance on Overlays.

Learn more: https://csrc.nist.gov/Projects/risk-management/sp800-53-controls/overlay-repository/overlay-overview

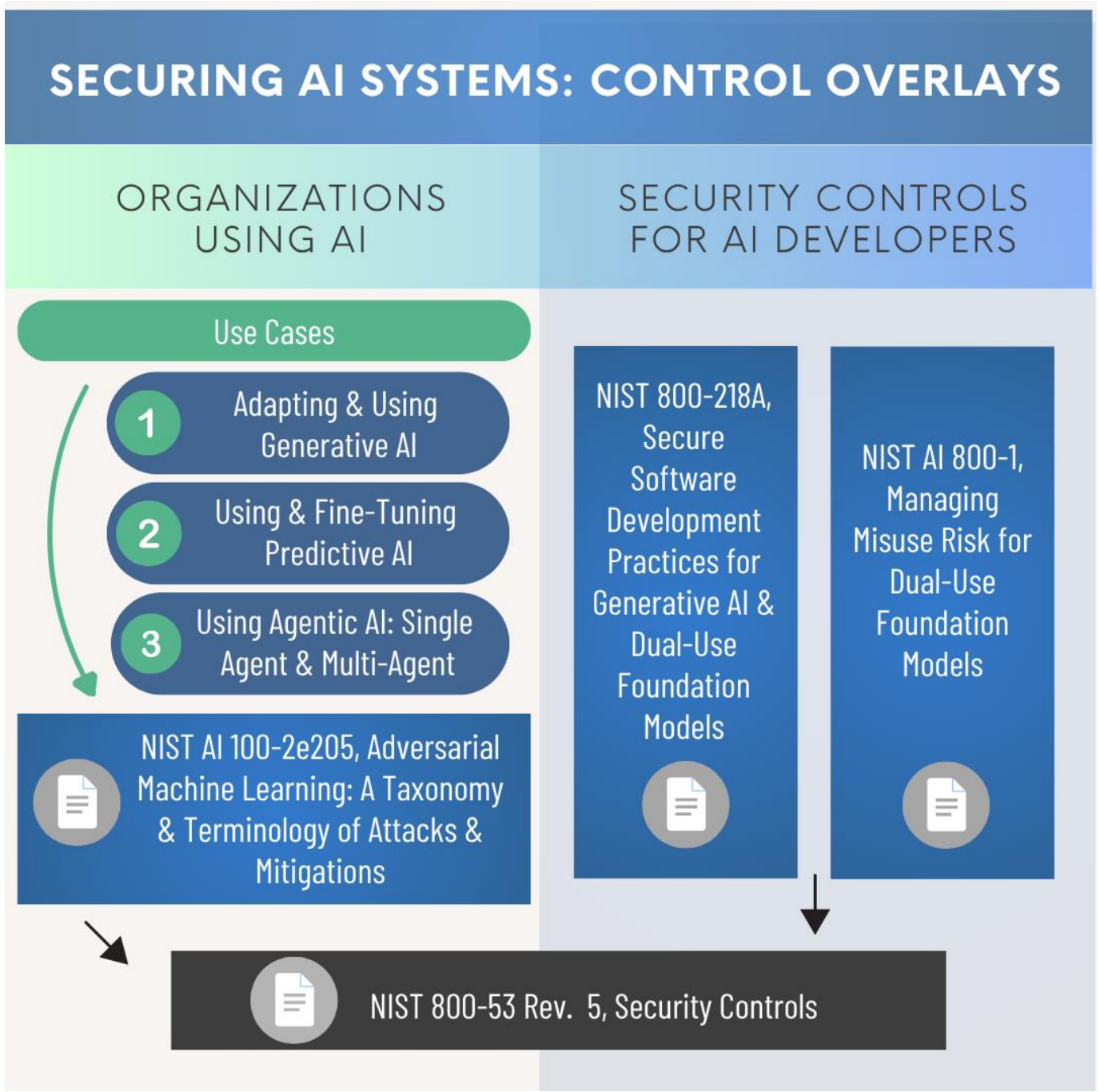
Proposed Methodology

This methodology assumes that the organization has:

- an organization-wide information security program in place to manage cybersecurity risk and includes a risk management strategy with explicit risk tolerance
- selected and implemented controls to manage risk (including governance)
- implemented a process to assess and monitor controls

Concept Paper Identifies Proposed Use Cases

Download the concept paper and learn more about the use cases:



Proposed Use Case 1: Using GenAI

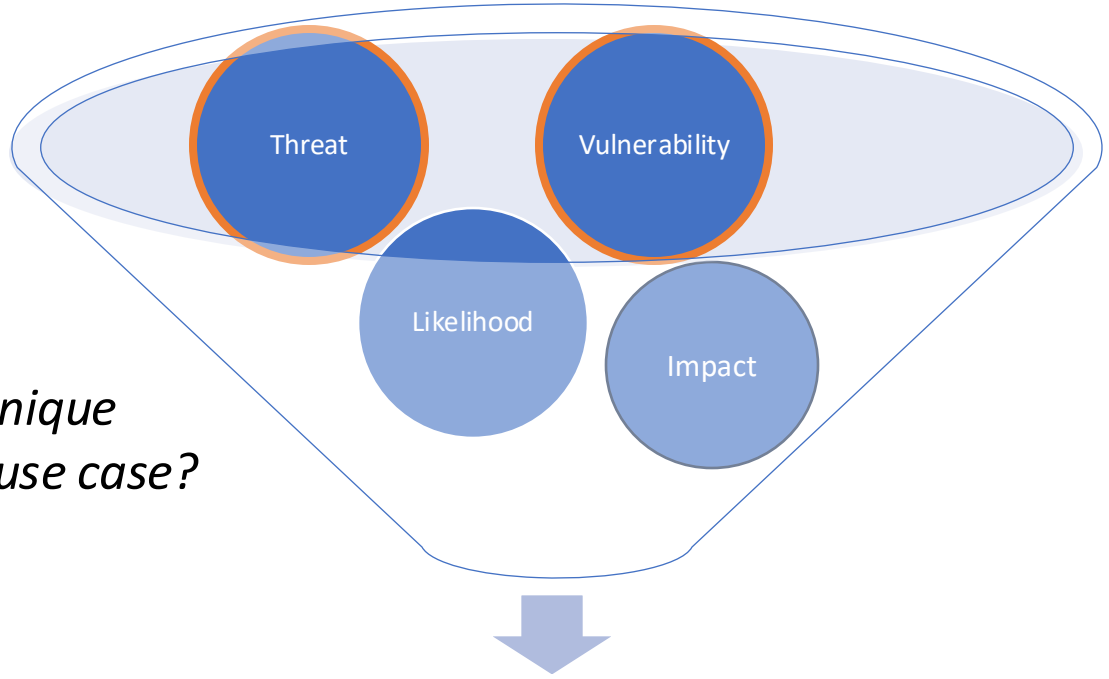
Audience	Organizations interested in using AI for creating new content (e.g., text, images).
Purpose	Generative AI creates new content (e.g., text, images, audio, video) based on user prompts by learning from large datasets and identifying patterns in the datasets
Use Case Description	This use case will cover examples of generative AI used for internal business augmentation (e.g., creating summaries, analyzing data) by internal users. A. On-premises hosted LLM with supporting infrastructure: Retrieval-augmented generation (RAG) with proprietary local data sources, mixed local and web sources, and external web sources only. B. B. Third-party hosted LLM with supporting infrastructure: RAG with proprietary local data sources, mixed local and web sources, and external web sources only.

See all of the use cases in the concept paper:

AI Cybersecurity Risk Assessment

**Note the risk assessment is based on the use case and includes the assumptions from slide 10*

What is unique to GenAI use case?



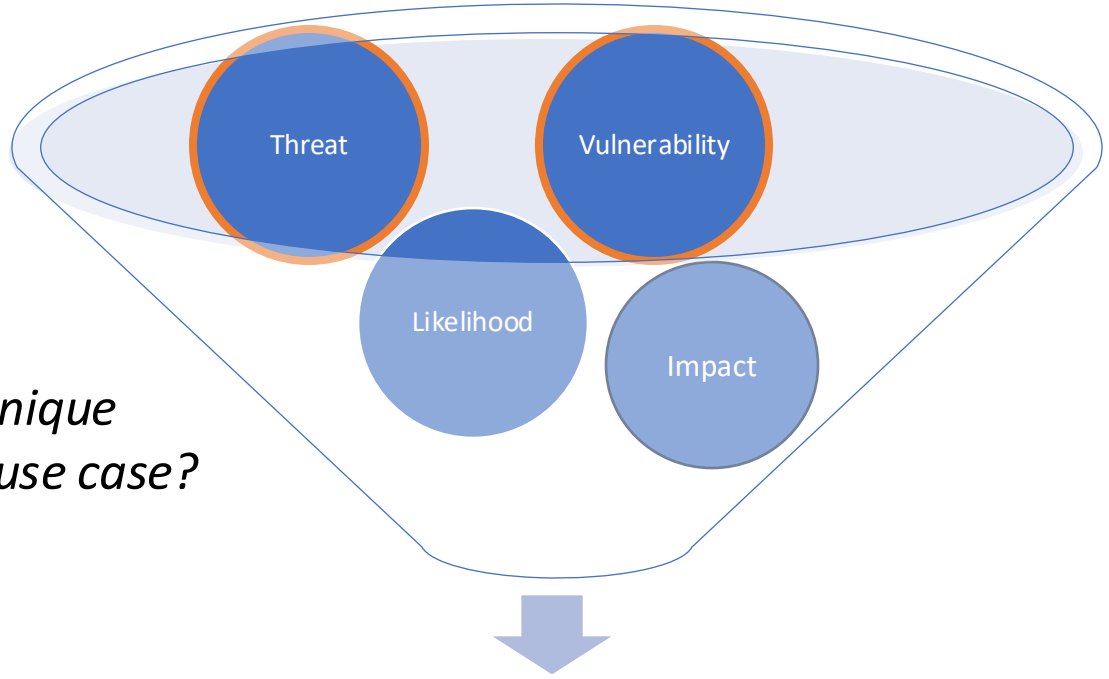
Examples

Data poisoning
Data extraction
Indirect prompt injection
Targeted Poisoning
Membership inference
Prompt injection
Backdoor poisoning
Prompt extraction
Misaligned Objects
Training data attacks
Leaking information from use interactions
Compromise connected resources
Model poisoning

AI Cybersecurity Risk Assessment

**Note the risk assessment is based on the use case and includes the assumptions from slide 10*

What is unique to GenAI use case?



Examples

Data poisoning
Data extraction
Indirect prompt injection
Targeted Poisoning
Membership inference
Prompt injection
Backdoor poisoning
Prompt extraction
Misaligned Objects
Training data attacks
Leaking information from use interactions
Compromise connected resources
Model poisoning

Identify & Tailor Controls

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Notional Example

Risk	Description	Potential Controls to Manage Risk	Tailoring
Model Poisoning	Intentional corruption of an AI model's training data to create incorrect behavior, introduce biases, create hidden backdoors, or degrade performance	AC-06, Least Privilege CM-02, Baseline Configuration CM-03, Configuration Change Control SC-28, Protection of Information at Rest SC-32, Information System Partitioning SI-03, Malicious Code Protection SI-04, System Monitoring SI-07, Software, Firmware & Information Integrity SI-07(01), SW, FW& Information Integrity | Integrity Checks SI-10, Information Input Validation SR-03, Supply Chain Controls and Processes ….	*modifying a control that includes:* • *applying scoping considerations,* • *selecting compensating controls,* • *assigning specific values to organization-defined control parameters,* • *providing additional specification information for implementation*

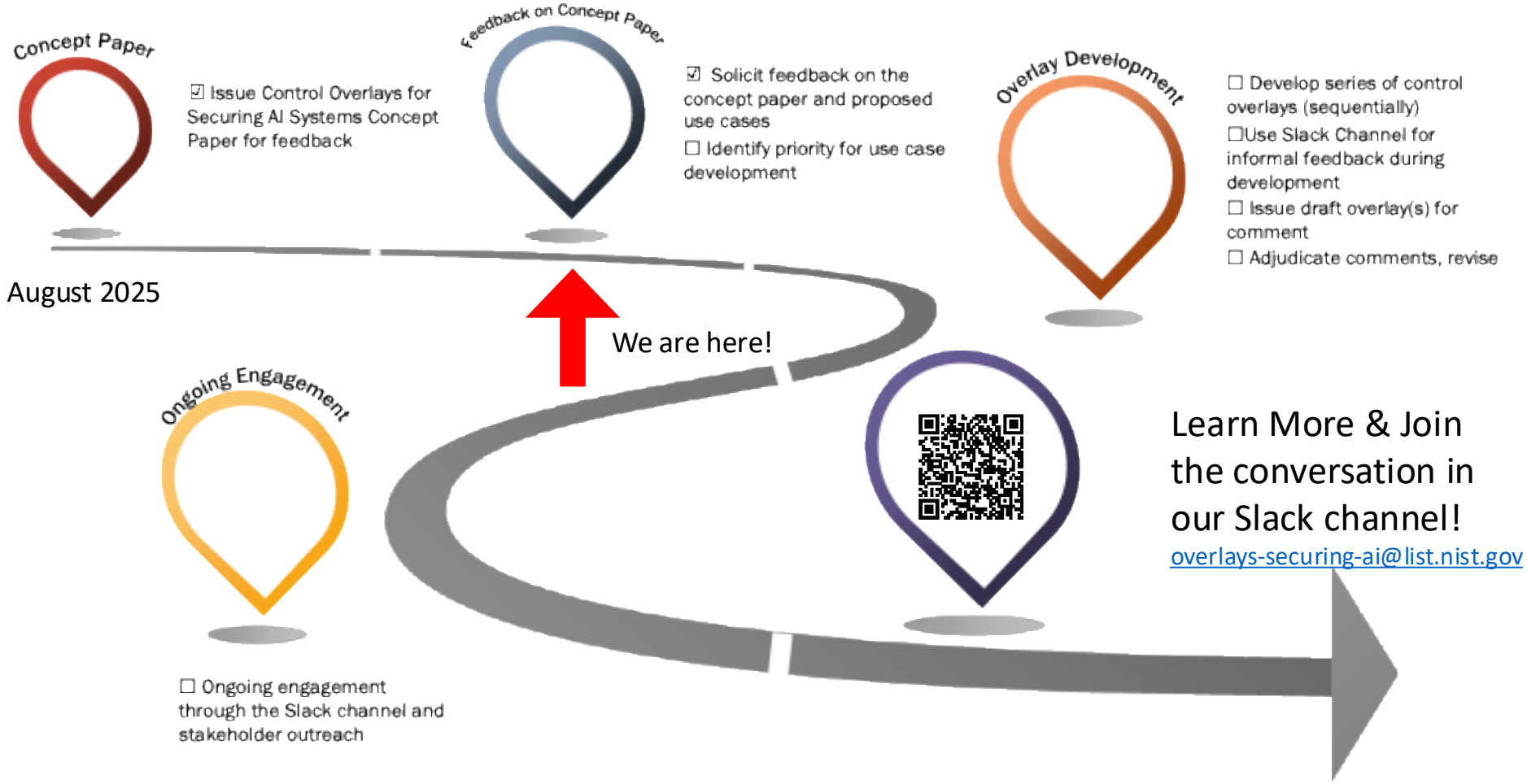
Gap Analysis

Notional Example

Risk	Potential Controls to Manage Risk	Organization Controls in Place
Model Poisoning	AC-06, Least Privilege CM-02, Baseline Configuration CM-03, Configuration Change Control SC-28, Protection of Information at Rest SC-32, Information System Partitioning SI-03, Malicious Code Protection SI-04, System Monitoring SI-07, Software, Firmware, and Information Integrity SI-07(01), Software, Firmware, and Information Integrity | Integrity Checks SI-10, Information Input Validation SR-03, Supply Chain Controls and Processes ….	- *Considers what controls are applied organization-wide and specific to the system* - *Not all controls are applied to all system components* - *Where are existing controls addressing the AI-specific risks?* - *What are existing controls NOT addressing AI-specific risks?*

Next Steps and How to Get Engaged

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE



Learn More & Join the conversation in our Slack channel!

overlays-securing-ai@list.nist.gov

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

STAY IN TOUCH

https://csrc.nist.gov/Projects/cosais

overlays-securing-AI@list.nist.gov