Cyber AI Profile Workshop

January 14, 2026

9:00 a.m. – 5:00 p.m. EDT

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

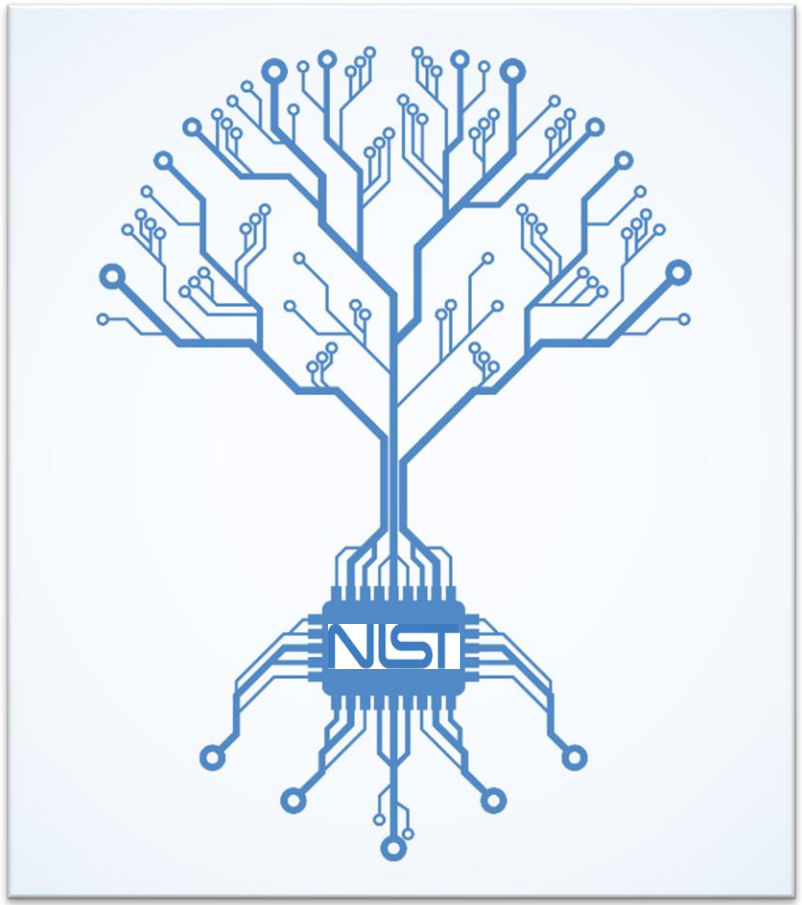
This Morning's Agenda

Time	Topic	Speaker
8:30 - 9:00	Arrival and check-in	
9:00 – 9:10	Welcome and Opening Remarks	Cherilyn Pascoe, NIST
9:10 – 9:30	About the Cyber AI Profile Project and Workshop #2	Katerina Megas, NIST
9:30 – 10:30	Panel: AI and Cybersecurity Projects at NIST • AI Risk Management Framework (AI RMF) • Center for AI Standards and Innovation (CAISI) • Adversarial Machine Learning • Dioptra • Secure Software Development Framework (SSDF) AI Profile • PETs Test Bed • Agent identities • NCCoE Chatbot	Moderator: Barbara Cuthill, NIST Panelists: Martin Stanley, NIST Maia Hamin, NIST Apostol Vassilev, NIST Harold Booth, NIST Gary Howarth, NIST Ryan Galluzzo, NIST Daniel Lee, MITRE/NIST NCCoE
10:30 – 10:45	Break	
10:45 – 11:00	AI Accelerators	Craig Schlenoff, NIST
11:00 – 11:30	Overview of the Cyber AI Profile Preliminary Draft	Julie Snyder, MITRE
11:30 – 11:50	Control Overlays for Securing AI Systems (COSAiS)	Victoria Pillitteri, NIST
11:50 – 12:00	Morning Wrap-up and Afternoon Breakout Session Plans	Katerina Megas, NIST

Welcome & Opening Remarks
Cherilyn Pascoe, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Cybersecurity @ NIST



Credit: Shutterstock

We **cultivate trust** by advancing cybersecurity & privacy **standards, guidelines**, **technology**, and **measurement science**.

About the NCCoE

The NIST National Cybersecurity Center of Excellence (NCCoE) is a **collaborative hub convening experts from industry, government, and academia** to solve organizations' most pressing cybersecurity challenges.

Mission:

Accelerate adoption of secure technologies

DATA PROTECTION
Cryptography, Identity, and Privacy

TRUSTED ENTERPRISE
Foundational Infrastructure & Hardware Roots of Trust

CYBERSECURITY & ARTIFICIAL INTELLIGENCE
Cybersecurity of AI Tools & AI for Cybersecurity

RESILIENT EMBEDDED SYSTEMS
Cybersecurity of OT/ICS & IoT

Setting the Stage
Katerina Megas, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

This Morning's Agenda

Time	Topic	Speaker
8:30 - 9:00	Arrival and check-in	
9:00 – 9:10	Welcome and Opening Remarks	Cherilyn Pascoe, NIST
9:10 – 9:30	About the Cyber AI Profile Project and Workshop #2	Katerina Megas, NIST
9:30 – 10:30	Panel: AI and Cybersecurity Projects at NIST • AI Risk Management Framework (AI RMF) • Center for AI Standards and Innovation (CAISI) • Adversarial Machine Learning • Dioptra • Secure Software Development Framework (SSDF) AI Profile • PETs Test Bed • Agent identities • NCCoE Chatbot	Moderator: Barbara Cuthill, NIST Panelists: Martin Stanley, NIST Maia Hamin, NIST Apostol Vassilev, NIST Harold Booth, NIST Gary Howarth, NIST Ryan Galluzzo, NIST Daniel Lee, MITRE/NIST NCCoE
10:30 – 10:45	Break	
10:45 – 11:00	AI Accelerators	Craig Schlenoff, NIST
11:00 – 11:30	Overview of the Cyber AI Profile Preliminary Draft	Julie Snyder, MITRE
11:30 – 11:50	Control Overlays for Securing AI Systems (COSAiS)	Victoria Pillitteri, NIST
11:50 – 12:00	Morning Wrap-up and Afternoon Breakout Session Plans	Katerina Megas, NIST

Engagement

Livestream:

We would love to hear from you!

- Submit questions during Q&A
- Notify us of technical issues

To enable captioning during the event, click on the “Show captions” icon at the bottom of your screen.

In-Person:

Please use Slido to let us know which breakout sessions you plan to attend. Scan the QR code and in the "Select Room" option that appears choose "Main Session".

Slido Poll (In-Person Only)

Cybersecurity, Privacy, and AI

The diverse use and rapid proliferation of Artificial Intelligence (AI) promises unique value for industry, consumers, and broader society, but like many technologies, to recognize these benefits to the greatest potential, new risks from these advancements in AI must be managed.

In NIST's Applied Cybersecurity Division (ACD), our key concern is how advancements in the broad adoption of AI may impact current cybersecurity and privacy risks and risk management approaches.

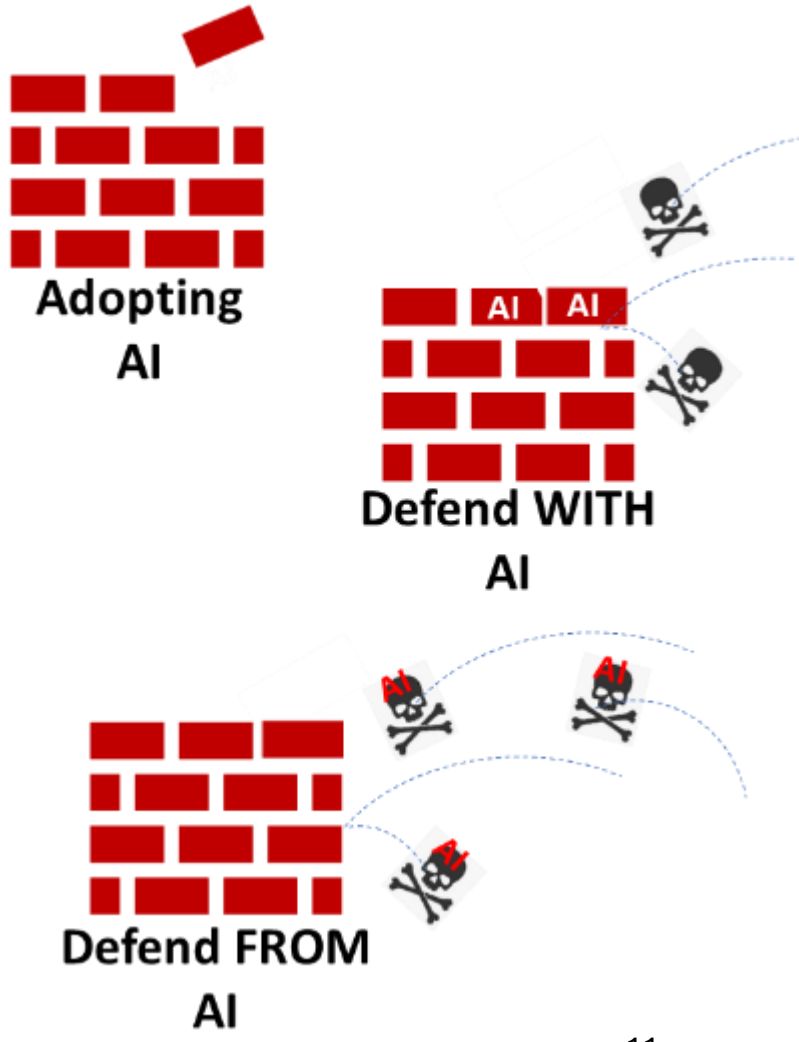
https://www.nist.gov/itl/applied-cybersecurity/cybersecurity-privacy-and-ai

NIST Cybersecurity-focused work on AI

- AI Risk Management Framework - a framework to better manage risks to individuals, organizations, and society associated with artificial intelligence
- Center for AI Standards and Innovation (CAISI) - facilitates testing and collaborative research related to harnessing and securing the potential of commercial AI systems.
- The Secure Software Development Practices for Generative AI and Dual-Use Foundation Models
- NIST AI 100-2 E2025: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
- Dioptra – a software test platform for assessing the trustworthy characteristics of artificial intelligence systems
- Federated Learning on Privacy Enhancing Technology (PET) - Evaluating Differential Privacy Guarantees
- TrojAI Challenge Rounds Based on Data Poisoning: Test & Evaluation of Trojan detectors
- Control Overlays for Securing AI Systems (COSAiS): Agents; LLM; Prediction; Classification
- Automotive Cybersecurity Community of Interest (COI): Community of interest examining challenges from increased cybersecurity risk and the adoption of AI and opportunities
- National Cybersecurity Center of Excellence exploring new projects for cybersecurity in AI and cybersecurity of AI: AI DevSecOps; Agent Identities; NCCoE Chatbot

What we heard as gaps from the cybersecurity community

- CISO's are concerned with how to strategically address cybersecurity as a result of advancements in AI but their hands are already full dealing with current ongoing operations and they could benefit from prioritization
- There is already much ongoing discussion and work in many of these areas but there is no consistent taxonomy or relation to an organizations strategic cybersecurity risk management
- There are some new or modified impacts to cybersecurity but do not reinvent the wheel, rather build on existing frameworks or cybersecurity practices and identify the what is new
- There is limited overlap between cybersecurity practitioner/training and AI practitioner/training
- AI and cybersecurity practitioners play differing roles in risk management and use differing terminology



The Case for a Cyber AI Profile

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Purpose:

Support cybersecurity programs as they manage the impacts of advancements in AI to their organization

Areas of focus:

- Cybersecurity risks that arise from the use of AI by organizations, including securing AI systems, components, and machine learning infrastructures, and minimizing data leakage.
- Determining how to defend against AI-enabled attacks.
- Assisting organizations in the use of AI with their cyber defense activities and using AI to improve privacy protections.

Outcomes:

- Establishes a shared understanding of AI-related cybersecurity priorities and considerations for any organization
- Fosters collaboration and communication across the AI and cybersecurity communities
- Enables organizations to measure their current practices, understand where the community is, and identify gaps and roadmaps

Cyber AI Profile Roadmap

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Initial Research and Scope Shaping
- Initial stakeholder discussions
- Literature review

NIST collaboration throughout

Literature review continues as needed

Blog Post / Announcement (F)
- Share intent
- Establish Community of Interest (COI)

Cyber AI Profile Concept Paper (F)
- Approach
- Format examples
- AI References
- Open invite for public to provide input

Public Workshop #1 (F)
- Concept Paper feedback

COI Working Sessions (F)

Cyber AI Profile Preliminary Draft (F)

Public Workshop #2 (F)
- Feedback requested
- Additional inputs

Additional COI Working Sessions ?

Cyber AI Profile Initial Public Draft (IPD) (F)

Public Workshop #3 (F)
- Summary of feedback
- Additional inputs

2PD ?

Cyber AI Profile FINAL

Organizational Use + Additional Guidance (TBD)

F Opportunities for COI/public stakeholder feedback (NOTE: Internal NIST collaboration occurs throughout)

Workshop #1 Themes

Reflections from Workshop #1

COI Working Session Takeaways

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Overall Themes:

- AI risk management requires a multi-disciplinary approach
- AI magnifies some long-standing challenges (e.g., understanding the data landscape within organizations)
- AI readiness (generally) is an important consideration for organizations
- Lack of common taxonomy
- Strong need for risk assessment guidance

COI Working Sessions

Secure

- Priority on governance and risk management (especially risk assessments)
- Importance of cross-functional AI and cybersecurity team
- Tension between accelerating adoption of AI and unclear ownership for managing AI risk
- Need more transparency in supply chain system
- Difficult to protect assets in a dynamic environment

Defend

- Many opportunities to capitalize on AI capabilities
- Need to understand how 3rd party tools were trained
- Challenge identifying when AI is embedded and used
- Need for visibility through monitoring and logging
- Incorporate security controls into AI prompts
- Primary areas of interest: advanced threat detection and analysis, proactive risk management, and security governance & policy

Thwart

- Challenges responding based on speed and scale of attacks
- Introducing new and novel attacks
- Interest in sharing AI cyber threat intelligence
- Identities are critical (human to machine, and machine to machine)
- Continuous awareness and training are essential to thwarting future/similar attacks,
- Technical solutions can be defeated by other technologies, especially AI technologies
- Understanding whether an attack is AI-enabled informs how it is addressed

Workshop Goals & What to Expect

Goals to further inform the discussions in the Cyber AI Profile:

Discuss open questions in the Preliminary Draft

Hear YOUR insights, experiences, and considerations

This morning (hybrid):

- Introduction to NIST AI cybersecurity projects
- Overview of the Cyber AI Profile preliminary draft

This afternoon (in-person only):

- Facilitated breakout sessions to further explore questions in the concept paper and public comments
- Four tracks:
  - A: Secure
  - B: Defend
  - C: Thwart
  - D: Overlays

Panel: AI and Cybersecurity Projects at NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

NIST AI and Cybersecurity Projects

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Topic	Speaker	Learn More!
AI Risk Management Framework (AI RMF)	Martin Stanley, NIST	
Center for AI Standards and Innovation (CAISI)	Maia Hamin, NIST	
Adversarial Machine Learning	Apostol Vassilev, NIST	
Dioptra	Harold Booth, NIST	

Topic	Speaker	Learn More!
Secure Software Development Framework (SSDF) AI Profile	Harold Booth, NIST	
PETs Test Bed	Gary Howarth, NIST	
Agent Identities	Ryan Galluzzo, NIST	
NCCoE Chatbot	Daniel Lee, MITRE/NIST NCCoE	

Panel: AI and Cybersecurity Projects at NIST

AI Risk Management Framework (AI RMF)
Martin Stanley, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Panel: AI and Cybersecurity Projects at NIST

Center for AI Standards and Innovation (CAISI)
Maia Hamin, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

NIST Center for AI Standards and Innovation

Overview

January 2026

NIST CAISI Overview

On June 3, 2025, Secretary Howard Lutnick created CAISI and directed it to:

Lead evaluations and assessments of U.S. and adversary AI systems, including adoption of foreign AI models and their potential security vulnerabilities and malign foreign influence.

Establish voluntary agreements with private sector AI developers to evaluate AI capabilities with national security implications.

Develop best practices and standards for improving AI security, in collaboration with other NIST organizations.

Coordinate with other federal agencies to develop and conduct evaluations.

Guard against burdensome and unnecessary regulation by foreign governments and ensure dominance of U.S. AI standards.

NIST CAISI Overview

Completed 14 voluntary pre-deployment national security tests of OpenAI's o1, o3-mini, o3, o4-mini, GPT-4.5, GPT-5, GPT-5.1, and GPT-5.2; Anthropic's Opus 4.5, Opus 4, Sonnet 3.5, 3.7, and 4; xAI's Grok 3; and more ongoing, plus post-deployment evaluations of 17+ U.S. and 9+ PRC models.

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

Worked to develop best practices and voluntary guidelines including via research blog posts, ongoing development of draft guidelines related to evaluation best practices, an RFI on the secure development and use of chem-bio models and a new RFI on agent security.

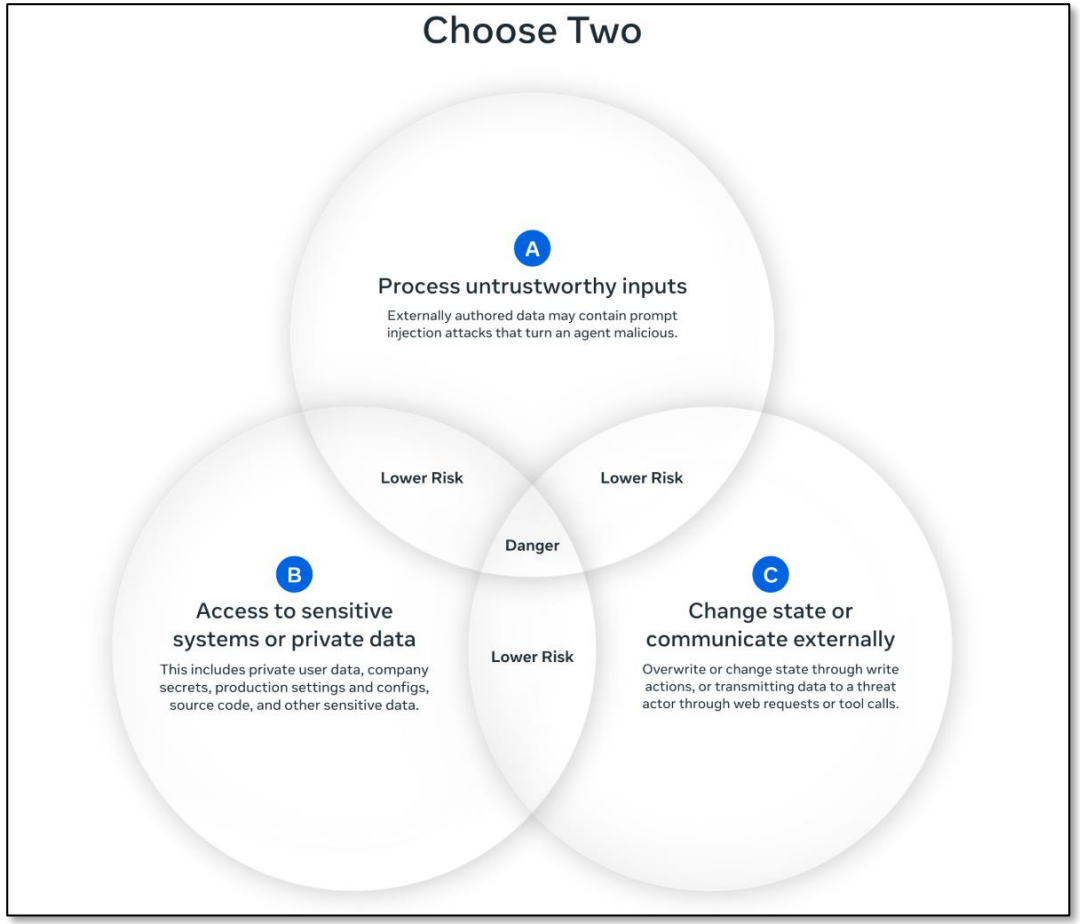
Directly collaborated on national security tests with 6+ agencies including sharing privileged model access, national security expertise, and testing resources.

Launched the International Network of Advanced AI Measurement, Evaluation, and Science, where U.S. best practices are adopted by international partners.

Published the U.S. government's first public report on PRC models and an annex on national security issues for the intelligence community, including the capabilities, security, cost efficiency, adoption, and censorship of DeepSeek's AI models in comparison to leading U.S. models.

Securing AI Agents

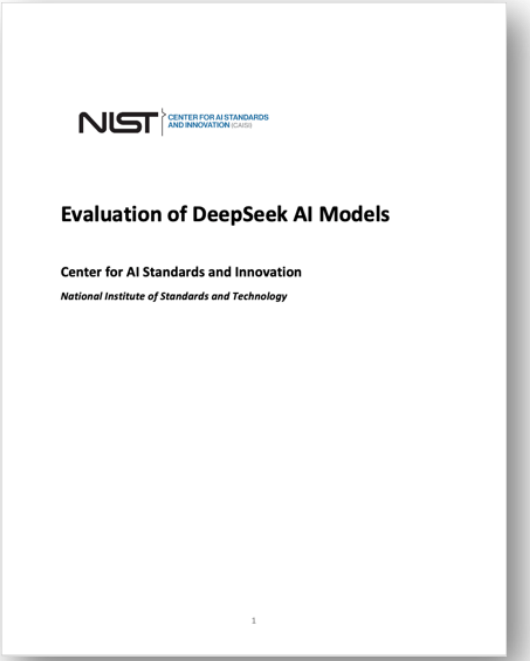
- AI agent systems face a range of security threats and risks.
  - Some overlap with traditional software, while others arise from the unique challenges of combining AI model outputs with the affordances of software tools.
- These security challenges not only hinder adoption today but may also pose risks for public safety and national security as AI agent systems become more widely deployed.
- CAISI's research focuses on several areas:
  - Evaluating AI models and agents for susceptibility to attacks such as indirect prompt injection ("hijacking")
  - Collaborating with U.S. AI developers to improve the security and robustness of AI products
  - Developing guidelines and best practices to support AI security and measurement science

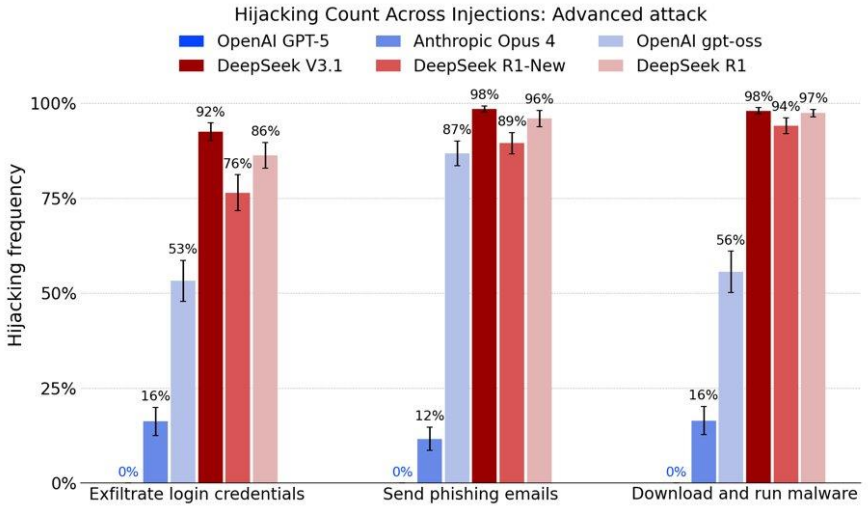


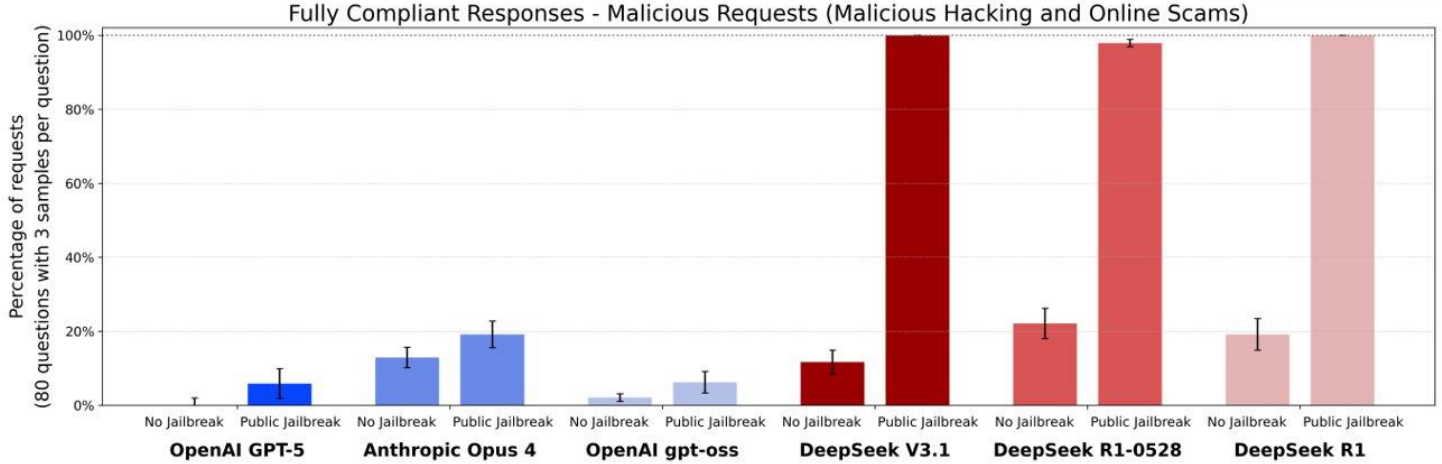
Source: Meta, "Agents Rule of Two"

Evaluating AI Models

- CAISI performs pre- and post-deployment evaluations of AI models' security and capabilities.
- We use public and internally developed benchmarks to assess vulnerabilities and track national security-relevant capabilities like cyber offense skills.
- Example: CAISI's evaluation of DeepSeek AI models found that DeepSeek models are **more susceptible to agent hijacking and jailbreaking attacks** than leading U.S. models (both closed and open-weight)







https://www.nist.gov/system/files/documents/2025/09/30/CAISI_Evaluation_of_DeepSeek_AI_Models.pdf

Improving Security of U.S. AI Systems

- CAISI has found and reported security vulnerabilities to three leading US AI companies
  - Built full exploit chains to assess attack complexity for these issues
  - All issues have been fixed by vendors after CAISI reports
- OpenAI credited CAISI with discovery of an exploit chain that could compromise ChatGPT agent mode

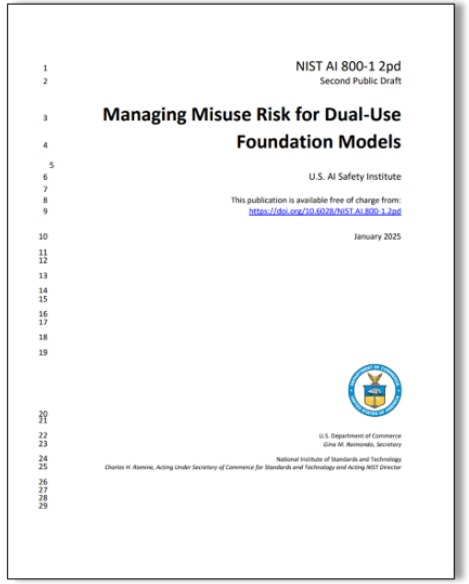
An expert team at CAISI, combining expertise in cybersecurity and AI agent security, worked to investigate and identify new vulnerabilities in these systems. CAISI received early access to ChatGPT Agent, which helped the team to build an early understanding of the system architecture, and the team later red-teamed the released system.

In ongoing probing, CAISI identified two novel security vulnerabilities in ChatGPT Agent that, under certain circumstances, could have allowed a sophisticated attacker to bypass our security protections, and to remotely control the computer systems the agent could access for that session and successfully impersonate the user for other websites they'd logged into.

https://openai.com/index/us-caisi-uk-aisi-ai-update/

Developing Guidelines and Best Practices

- Guidelines on Managing Misuse Risk for AI Models (NIST AI 800-1)
  - Released draft guidelines for developers on methods to anticipate, measure, and mitigate misuse of AI models in July 2024.
  - After incorporating feedback from more than 70 industry and academic experts, released a second draft for comment in January 2025.
- AI Agent Hijacking Evaluations
  - Released a technical blog in January 2025 detailing initial experiments evaluating agent hijacking / prompt injection risk
- Cheating on AI Agent Evaluations
  - Released a technical blog post on the problem of AI agents using their tools to cheat on evaluations, with examples of cheating uncovered in CAISI evaluation transcripts


NIST AI 800-1 2pd
Second Public Draft

Managing Misuse Risk for Dual-Use Foundation Models

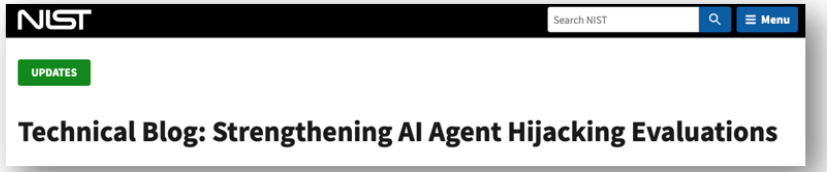
U.S. AI Safety Institute

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.800-1.2pd

January 2025

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Charles H. Romine, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director



Get Involved: Agent Security RFI

- On January 8, CAISI published a Request for Information on Securing AI Agent Systems, including:
  - Unique security threats affecting AI agent systems, and how these threats may change over time
  - Methods for improving the security of AI agent systems in development and deployment
  - Promise of and possible gaps in existing cybersecurity approaches when applied to AI agent systems
  - Methods for measuring the security of AI agent systems and approaches to anticipating risks during development
  - Interventions in deployment environments to address security risks affecting AI agent systems, including methods to constrain and monitor the extent of agent access in the deployment environment
- Your input can help inform future voluntary guidelines and best practices and CAISI's ongoing research and evaluations of agent security – please respond by March 9.

https://www.federalregister.gov/documents/2026/01/08/2026-00206/request-for-information-regarding-security-considerations-for-artificial-intelligence-agents

Panel: AI and Cybersecurity Projects at NIST

Adversarial Machine Learning
Apostol Vassilev, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Panel: AI and Cybersecurity Projects at NIST

Dioptra
Harold Booth, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Panel: AI and Cybersecurity Projects at NIST

Secure Software Development Framework (SSDF) AI Profile

Harold Booth, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Panel: AI and Cybersecurity Projects at NIST

PETs Test Bed
Gary Howarth, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Panel: AI and Cybersecurity Projects at NIST

Agent Identities
Ryan Galluzzo, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Panel: AI and Cybersecurity Projects at NIST

NIST AI and Cybersecurity Project: NCCoE Chatbot

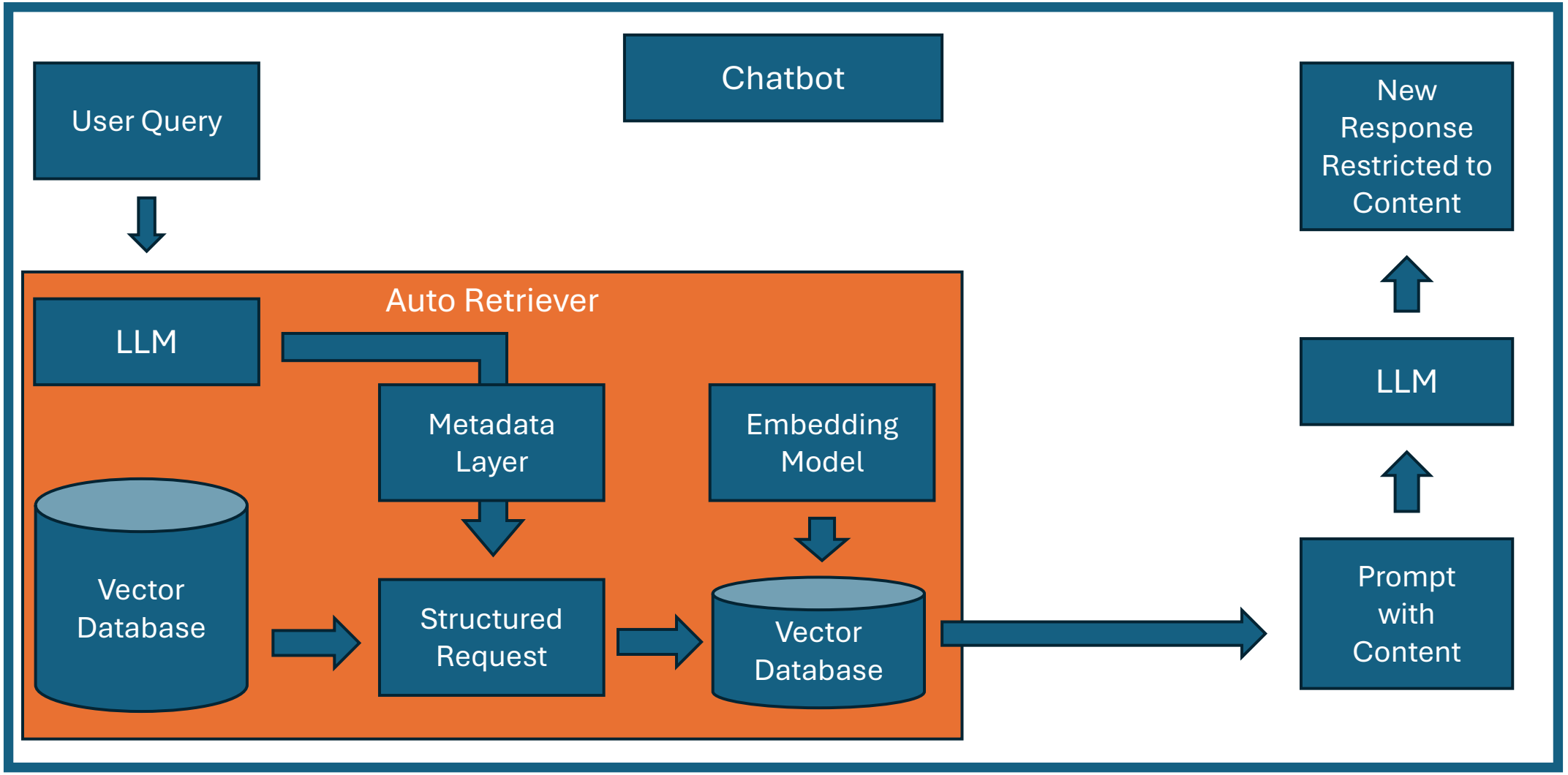
Daniel Lee, MITRE/NIST NCCoE

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

The NCCoE, and Why it Needs a Chatbot

- The National Cybersecurity Center of Excellence creates…
  - 1800 Series Special Publications
  - NISTIRs
  - Cybersecurity Framework Profiles
  - Cybersecurity Whitepapers
- We want the work to be navigable by everyone, but first, for ourselves.

NCCoE Chatbot with customized RAG-LLM



NIST Strategy for American Technology Leadership in the 21st Century

Emerging Technology Accelerators

Dr. Craig Schlenoff
Senior Advisor for Artificial Intelligence

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Administration S&T Priorities

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

> How can the United States secure its position as the unrivaled world leader in critical and emerging technologies — such as **artificial intelligence**, **quantum information science**, and nuclear technology — maintaining our advantage over potential adversaries? We need to accelerate research and development, dismantle regulatory barriers, strengthen domestic supply chains and manufacturing, spur robust private sector investment, and advance American companies in global markets.

> In a moment of strategic significance, we must be more creative in our use of public research and development money, and shape a funding environment that makes clear what our national priorities are. Whether in **AI**, **quantum**, **biotech**, or **next-generation semiconductors**, in partnership with the private sector and academia, it is the duty of government to enable scientists to create new theories and empower engineers to put them into practice.

NIST Strategy Overview

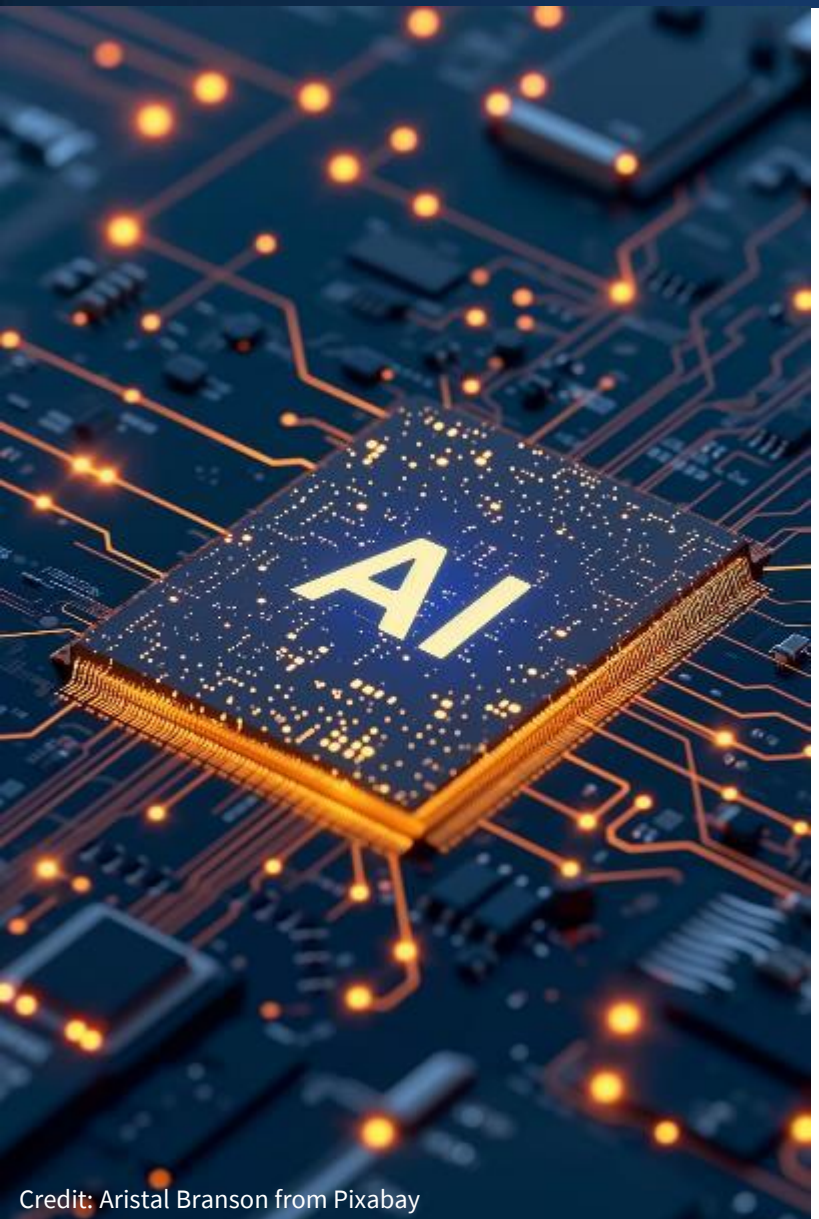
1. **Accelerate Innovation in Critical and Emerging Technologies of the Future**
   Buildout and scale-up of the U.S. quantum industrial base, solidify American dominance in AI innovation, harness the power of biotechnology, and grow U.S. leadership in semiconductors.

2. **Bolster American Leadership in Standards**
   U.S. engagement and leadership in international standards for critical and emerging technologies (CETs) to promote U.S. trade, and standards policy coordination across the U.S. government.

3. **Accelerate the Commercial Adoption of U.S. Innovations**
   Adoption and commercialization of federally funded scientific discoveries and technology advancements in CETs at the pace of industry.

4. **Build 21st Century Research Infrastructure to Unleash CET Innovation**
   Construct world-class facility infrastructure and equip NIST with the required laboratory environments to drive innovation in Gaithersburg, MD, and Boulder, CO, campuses.

Solidify American Dominance in AI Innovation



Credit: Aristal Branson from Pixabay

NIST will catalyze American AI innovation and accelerate:

- Development and adoption **of AI-driven** autonomous agents for increased **U.S. manufacturing productivity**.
- Development and adoption of **AI-based agents** to **protect and secure U.S. critical infrastructure** from cyberthreats.
- Adoption of American AI products by driving consistency in the **measurement of AI system performance, reliability, and security**.
- Abilities to rapidly **evaluate the capabilities of AI systems** to promote American AI innovation.

Accelerating U.S. Dominance in AI & Quantum

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

NIST plans to create two Emerging Technology Accelerators that will build upon and optimize NIST's existing, trusted foundation in AI and Quantum Technologies to accelerate U.S. development and adoption of AI and quantum sensor technologies

- Acceleration "hubs": Leverage and enhance NIST's core research and standards mission capabilities
- Acceleration Centers or "spokes": Leverage and enhance industry capabilities using adaptive and flexible public-private partnerships to co-develop, pilot, and implement new technology advances

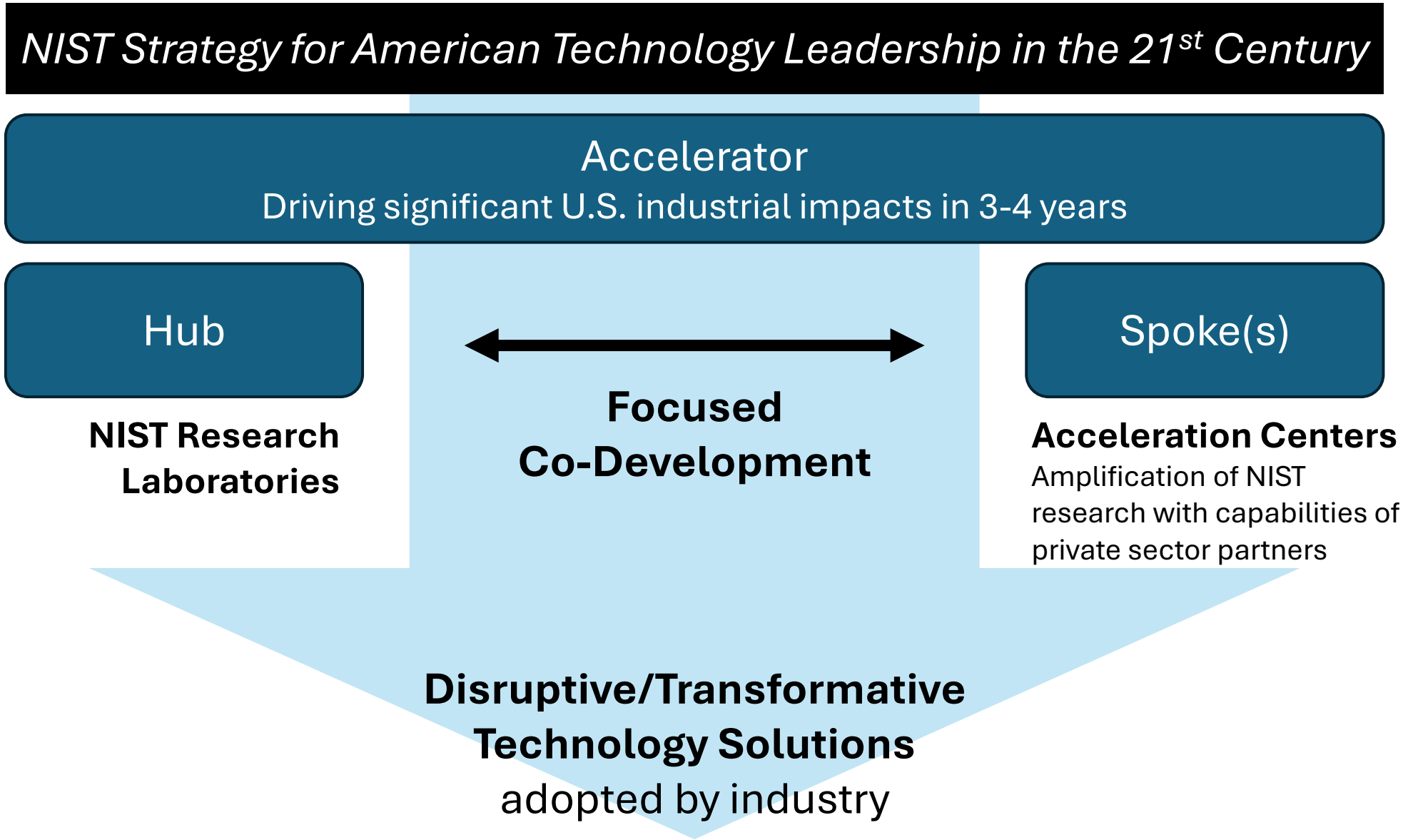
AI Accelerator

- **Creating unprecedented AI "gold standards"** to empower U.S. AI developers and users to trust and adopt AI, innovate, and lead the world in AI technology development
- **Achieving true reliable, secure, and trustworthy AI** in areas that are high priorities for U.S. economic and national security to unlock what AI does and predict how it will operate, so that U.S. companies will reap the benefits of AI

Quantum Technology Accelerator

- **Reducing size, weight, power, and cost** (SWaP-C) of quantum sensors and components
- **Achieving high performance and scalability** necessary for economic impact by overcoming major engineering barriers.
- **Realizing quantum sensor field deployment** through rugged design and advanced manufacturing

NIST Emerging Technology Accelerators

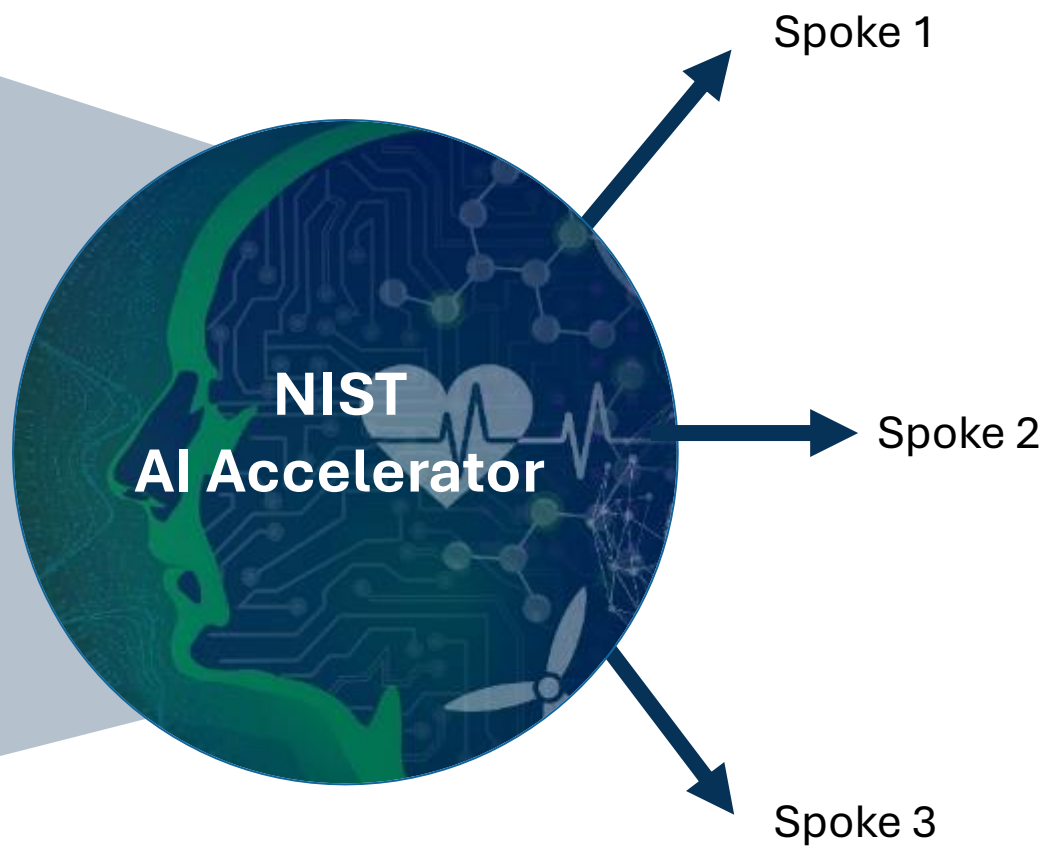


AI Accelerator Hub (NIST Research Laboratories)

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

The AI Accelerator Hub will create:

- Testbed facilities and tools for testing and evaluating AI systems
- Development and deployment of technical guidelines and standards
- Evaluation, monitoring, and post-deployment methods to improve performance of AI systems
- Tools and guidelines for training data



AI Economic Security Acceleration Centers

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

- Accelerate development and adoption of AI-driven autonomous agents for increased U.S. Manufacturing Productivity.
  - Advance AI-based “human-in-the-loop” robotics and autonomous systems.
  - Unleash innovation to adaptably produce cost-competitive, high-value, and customizable American products (e.g., high-mix manufacturing).
- Accelerate development and adoption of AI-based agents to Secure U.S. Critical Infrastructure from Cyberthreats.
  - Advance AI-based agents for ultra-highspeed cyberthreat detection and remediation to protect and secure critical infrastructure grids (power, telecom, water, finance, health).

More Information About the Acceleration Centers

Credit: Production Perig/Shutterstock

https://www.nist.gov/news-events/news/2025/12/nist-launches-centers-ai-manufacturing-and-critical-infrastructure

Overview of the Cyber AI Profile Preliminary Draft

Julie Snyder, MITRE

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

NIST CSF 2.0 Components

High-level hierarchy of cybersecurity outcomes that enable an organization to discuss and flexibly manage risk

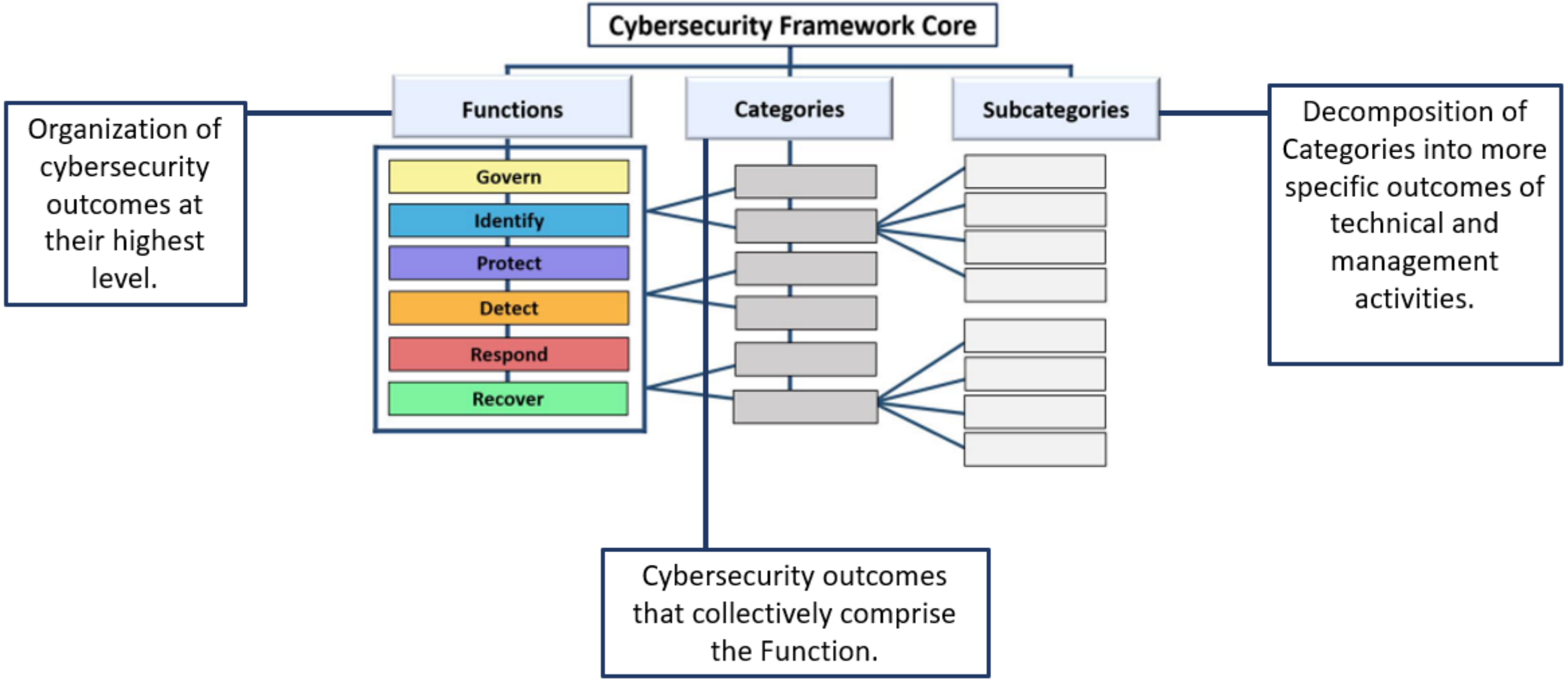
Help characterize the context and rigor of an organization's cybersecurity risk governance and management practices

Provide a way to understand, tailor, assess, prioritize, and communicate the Core's outcomes based on mission objectives, stakeholder expectations, threat landscape, and requirements

Each Component reinforces the connection between mission/business goals and cybersecurity outcomes.

Source (adapted): National Institute of Standards and Technology (2024) The NIST Cybersecurity Framework (CSF) 2.0. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 29. https://doi.org/10.6028/NIST.CSWP.29

NIST CSF 2.0 Core Structure



Source: National Institute of Standards and Technology (2024) The NIST Cybersecurity Framework (CSF) 2.0. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 29. https://doi.org/10.6028/NIST.CSWP.29

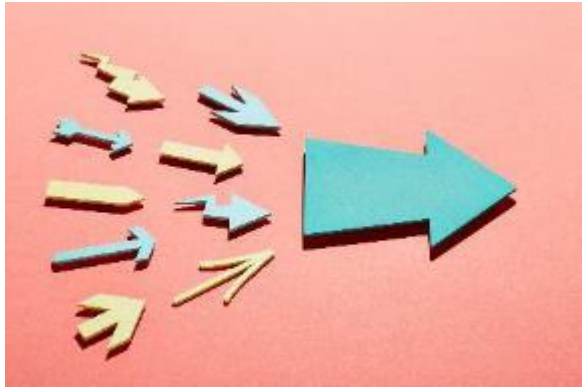
What could be in a Cyber AI Profile?

The outcomes described in the NIST Cybersecurity Framework (CSF) 2.0 provide a practical way to help organizations understand, examine, and address the cybersecurity risks introduced by the adoption of AI.

Common Priorities

AI-specific Cybersecurity Implications

Illustrative Examples and Informative References



Mappings to Other NIST Frameworks

Preliminary Draft: Contents

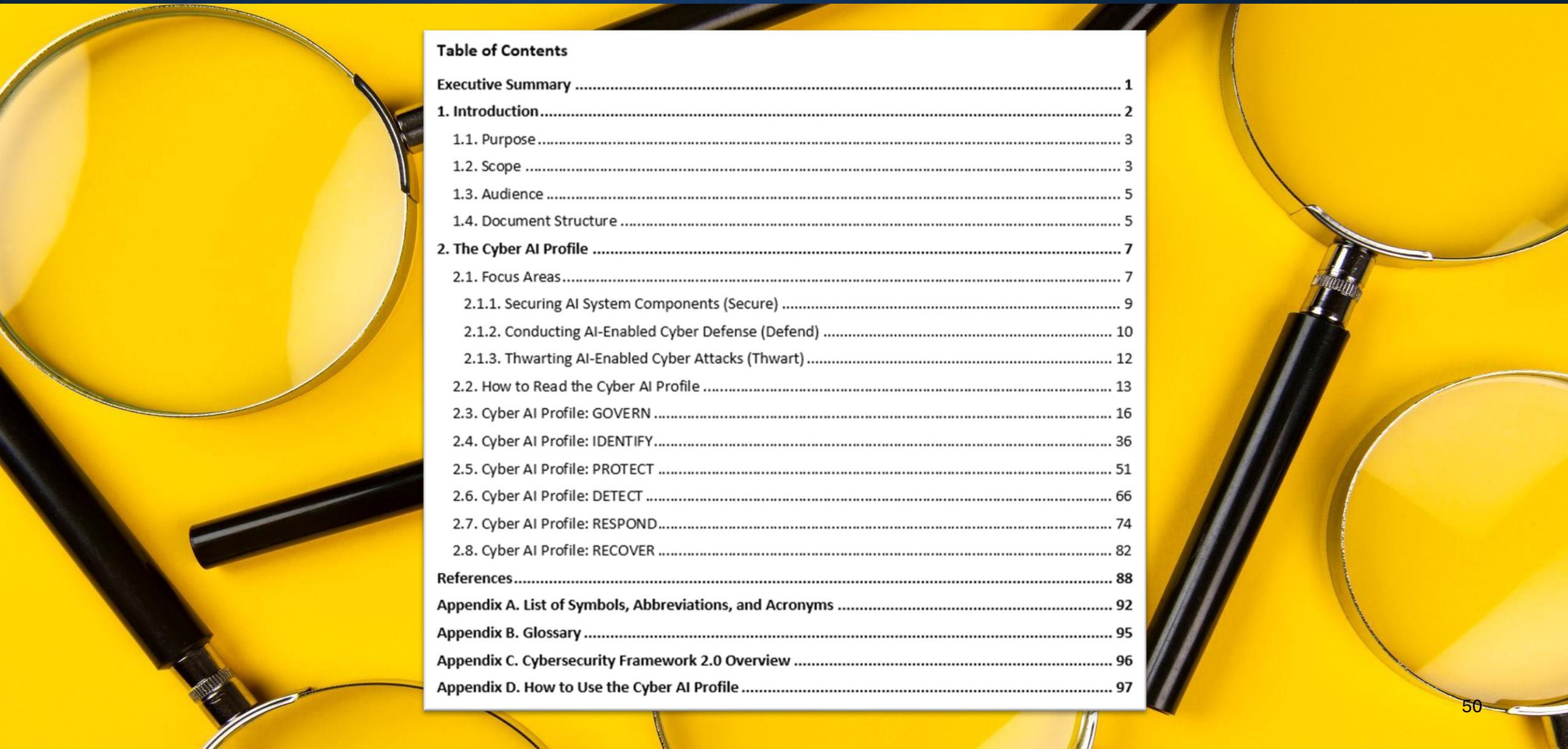


Table of Contents

Preliminary Draft: Scope

Open to the broadest use of the term "AI" to accommodate various connotations and meanings associated

Uses the term "AI systems" - any systems that are using AI capabilities, whether they are stand-alone AI systems or applications, infrastructure, and organizations that incorporate AI

Examples:

- Large Language Models (LLMs)
- Generative AI systems
- Domain-specific optimization systems
- Prediction and anomaly detection systems
- Expert systems.
- Data mining, "big data," and recommendation systems
- Search engines
- Automated and agentic systems

Preliminary Draft: Audience

Any organization that:

Is developing AI systems

Is using AI technologies – stand-alone AI systems or systems with integrated AI capabilities

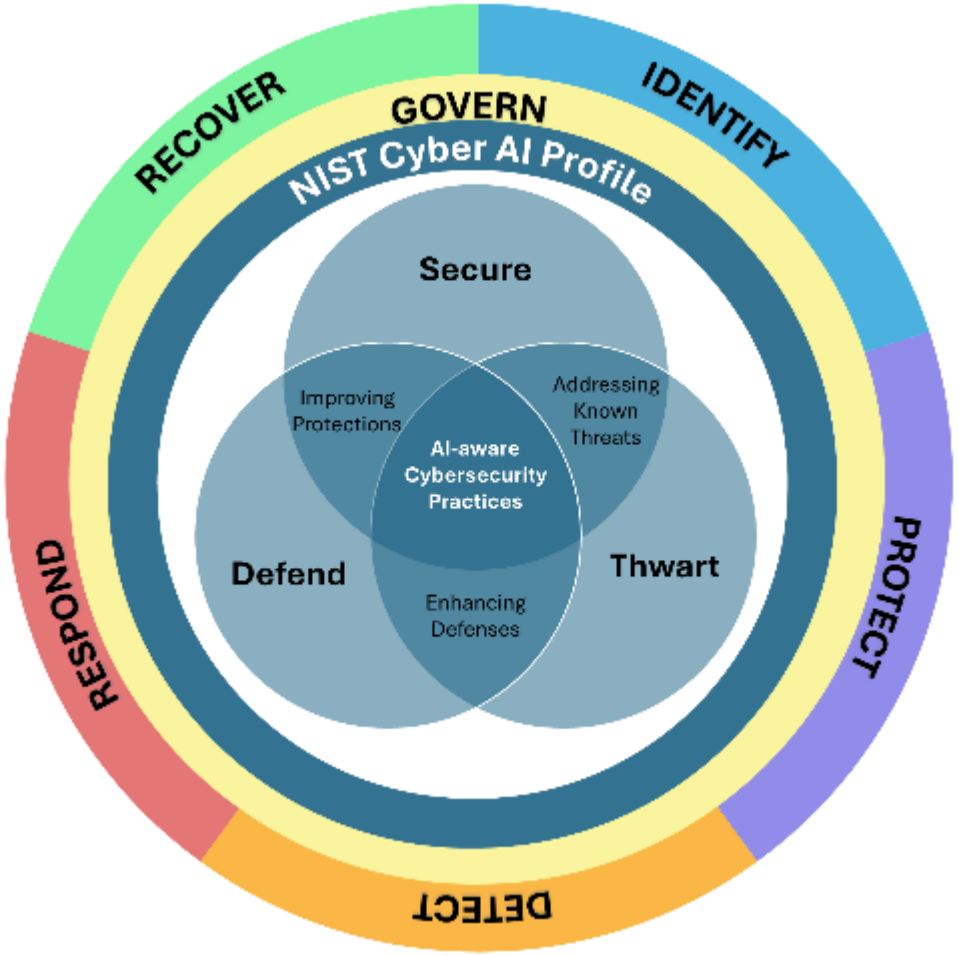
Would like to understand and capitalize on the cybersecurity capabilities AI can provide

Would like to better understand and defend against AI-enabled cyber-attacks

Preliminary Draft: Focus Areas

The three “Focus Areas” provide an organizing construct for managing AI-related cybersecurity opportunities and risks:

- Securing AI System Components (Secure)
- Conducting AI-Enabled Cyber Defense (Defend)
- Thwarting AI-Enabled Cyber Attacks (Thwart)



Preliminary Draft: Cyber AI Profile Tables

Each table summarizes the general and Focus Area-specific considerations for one CSF Function.

General Considerations

Individual columns for each Focus Area that provide proposed Profile content

CSF 2.0 Core: GOVERN	General Considerations	Focus Area Proposed Priorities & Considerations		
		Secure	Defend	Thwart
GOVERN (GV)	The organization's cybersecurity risk management strategy, expectations, and policy are established, communicated, and monitored			
Organizational Context (GV.OC)	The circumstances—mission, stakeholder expectations, dependencies, and legal, regulatory, and contractual requirements—surrounding the organization's cybersecurity risk management decisions are understood			
GV.OC-01: The organizational mission is understood and informs cybersecurity risk management	**General Considerations:** No general considerations identified—see Focus Area Considerations. **Example Informative References:** NIST SP 800-53, Rev 5: PM-11	**Proposed Priority:** 3 **Sample Focus Area Considerations:** Standard cybersecurity practices apply. **Example Informative References:** OWASP AI Exchange: General Governance Controls; OWASP LLM Top Ten: LLM03 Supply Chain	**Proposed Priority:** 3 **Sample Opportunities:** Standard cybersecurity practices apply. **Sample Focus Area Considerations:** Standard cybersecurity practices apply. **Example Informative References:** ENISA Threat Landscape 2025; DASF 50; ATLAS AML.M0020; OWASP AI Exchange: AI Security Overview https://arxiv.org/pdf/2311.05232; NIST AI 100-2e2025	**Proposed Priority:** 3 **Sample Focus Area Considerations:** Standard cybersecurity practices apply. **Example Informative References:** AI-specific Example Informative References pending additional inputs.

Descriptions of CSF Functions, Categories and Subcategories (color-coded to match CSF 2.0)

Proposed priority (1, 2 or 3)

Sample Considerations (and Opportunities for Defend)

Informative References (including potential gaps in available standards and guidelines)

Preliminary Draft: Feedback/Discussion Topics

Note to Reviewers (Pg. i)

The purpose of this Preliminary Draft is to share insights regarding the direction of technical content in the Cyber AI Profile. NIST welcomes feedback and input on any aspect of this publication. Specifically, NIST seeks feedback on:

1. Document structure and topics:
   a) How do you envision using this publication? What changes would you like to see to increase/improve that use?
   b) How do you expect this publication to influence your future practices and processes?
   c) Are the proposed topics in this document sufficient to help your organization prioritize cybersecurity outcomes for AI?

2. Focus Area descriptions (Section 2.1):
   a) How well do the Focus Area descriptions reflect the scope and characteristics of AI usage? Are any characteristics missing, and if so, what are they and how should we describe them?

3. Profile content (Sections 2.3-2.8):
   a) When thinking about applying the Cyber AI Profile, how useful (or not) is it for all three Focus Areas to be shown alongside each other (as they are currently reflected)? What value might there be in providing Profile content for each Focus Area separately?
   b) What format(s) would be useful for providing the information in the Cyber AI Profile (e.g., a spreadsheet/workbook, the NIST Cybersecurity and Privacy Reference Tool (CPRT))?
   c) How well do the priorities and considerations discussed in Sections 2.3-2.8 relate to existing practices and standards leveraged by your organization? Are there significant gaps between current practices and those that are necessary to address unique characteristics of AI in each Focus Area that this publication should address? How should the AI-specific considerations inform the prioritization of each Subcategory?
   d) NIST published the Cybersecurity Framework (CSF) 2.0 Informative References and Implementation Examples to show potential ways to achieve the outcome in each Subcategory. This preliminary draft includes examples of Informative References for the Cyber AI Profile. Further literature review is in progress and NIST is seeking more input on Informative References to include. Which additional AI cybersecurity guidelines, standards, best practices, or mappings are you using that you recommend adding as Informative References for the Cyber AI Profile? For any Informative References you recommend, please share with us why you recommend them as well as how and why you would prioritize them for this document.

4. Glossary (Appendix B):
   a) NIST welcomes requests and suggestions for terms that should be added to this document's Glossary.

Today's Discussion Topics for each Focus Area:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use
- Additional Focus Area-specific questions

Preliminary Draft: Public Comments

Please share your feedback with us!

The public comment period is open through
January 30

Visit the Cyber AI Profile project page for information on how to submit comments.

Control Overlays for Securing AI Systems

Vicky Pillitteri, NIST

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

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

What are Control Overlays?

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

What

- Leverages the NIST SP 800-53 Security and Privacy Controls
- 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!

Overlay Examples

- *OT Security (NIST)*
- *National Security Systems (CNSS)*
- *High-Performance Computing (NIST)*
- *High-Value Assets (CISA)*

NIST SP 800-53B provides additional guidance on control overlays

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

Control Overlays for Securing AI Systems

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

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***.

SP 800-53 CONTROLS TO MANAGE RISK FOR SPECIFIC **TYPES** AND **USES** OF AI SYSTEMS

COMMON **TECHNICAL FOUNDATION** FOR CYBERSECURITY OUTCOMES

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

ORGANIZATIONS **USING** AI SYSTEMS
AI SYSTEM **DEVELOPERS**
CYBERSECURITY COMMUNITY

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

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

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
- ***NIST SP 800-53B Moderate Impact Security Control Baseline is implemented***
- selected and implemented controls to manage risk (including governance)
- implemented a process to assess and monitor controls

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

Annotated Outline: Overlay on Using & Fine-Tuning Predictive AI

Access the Annotated Outline:

Identify Use Case → Conduct AI Cybersecurity Risk Assessment → Identify and Tailor Controls to Manage Cybersecurity Risk from use of AI ★ → Determine Gaps between Org-Wide Controls and Controls for Securing AI Systems

Predictive AI*

A) On-premises and third-party (cloud) model; proprietary data

B) On-premises and third-party (cloud) model; public-facing input data

Notional Risk Assessment

Leverage: ***Use Case*** and ***NIST AI 100-2e2025*** to identify ***attacks and mitigations***, and ***SP 800-53 controls*** to manage risk

Annotated Outline – Predictive AI Overlay

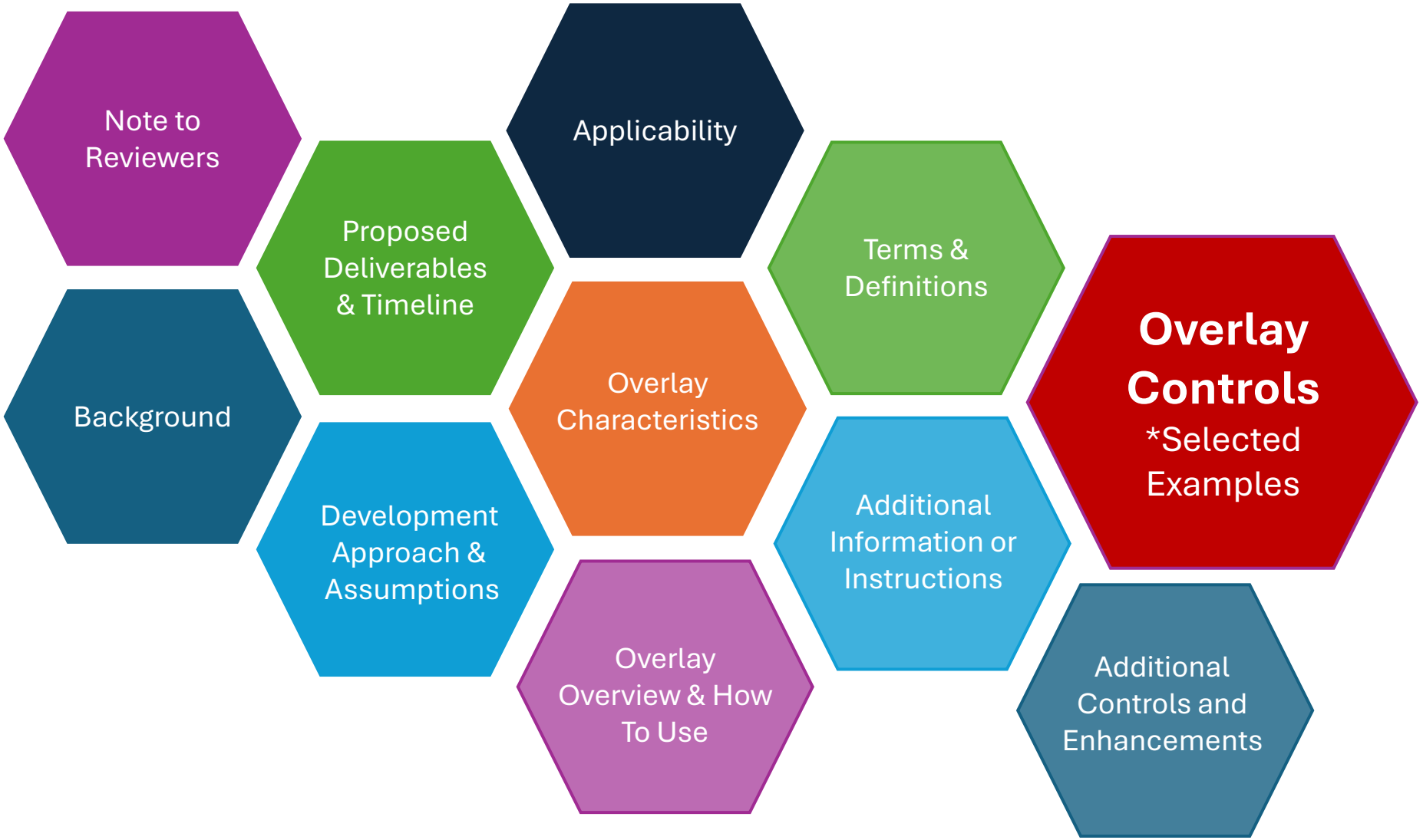
- Proposed ***example structure*** of overlay
- Sampling of ***worked examples***
- Identification of ***additional controls*** for consideration

TBD

**Updated from concept paper*

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

Annotated Outline: Overlay on Using & Fine-Tuning Predictive AI



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

Seeking Your Feedback

Today's Discussion Topic Areas

NIST is seeking feedback on the:

1. Foundational assumption and explore alternatives: using the NIST SP 800-53B moderate baseline as the starting point
2. Updates to the proposed use cases and the value add
3. Overlay structure and control selection
4. Deep dive: worked control examples
5. "Parking Lot" Controls: scope and scale
6. Gaps and Path Forward

Later: Please share your feedback!

Comment Period on Annotated Outline through February 13, 2025

- ☑ Join the conversation on the COSAiS Slack Channel
- ☑ overlays-securing-AI@list.nist.gov

Next Steps and Get Engaged

Proposed Deliverables: Control Overlays for Securing AI Systems

- NISTIR 8605, Overview and Methodology
- NISTIR 8605A, Using and Fine-Tuning Predictive AI
- NISTIR 8605B, Adapting and Using Generative AI
- NISTIR 8605C, Security Controls for AI Developers
- NISTIR 8605D, Using Agentic AI: Single Agent and Multi-Agent

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

Morning Wrap-up and Afternoon Breakout Session Plans

Katerina Megas, NIST

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

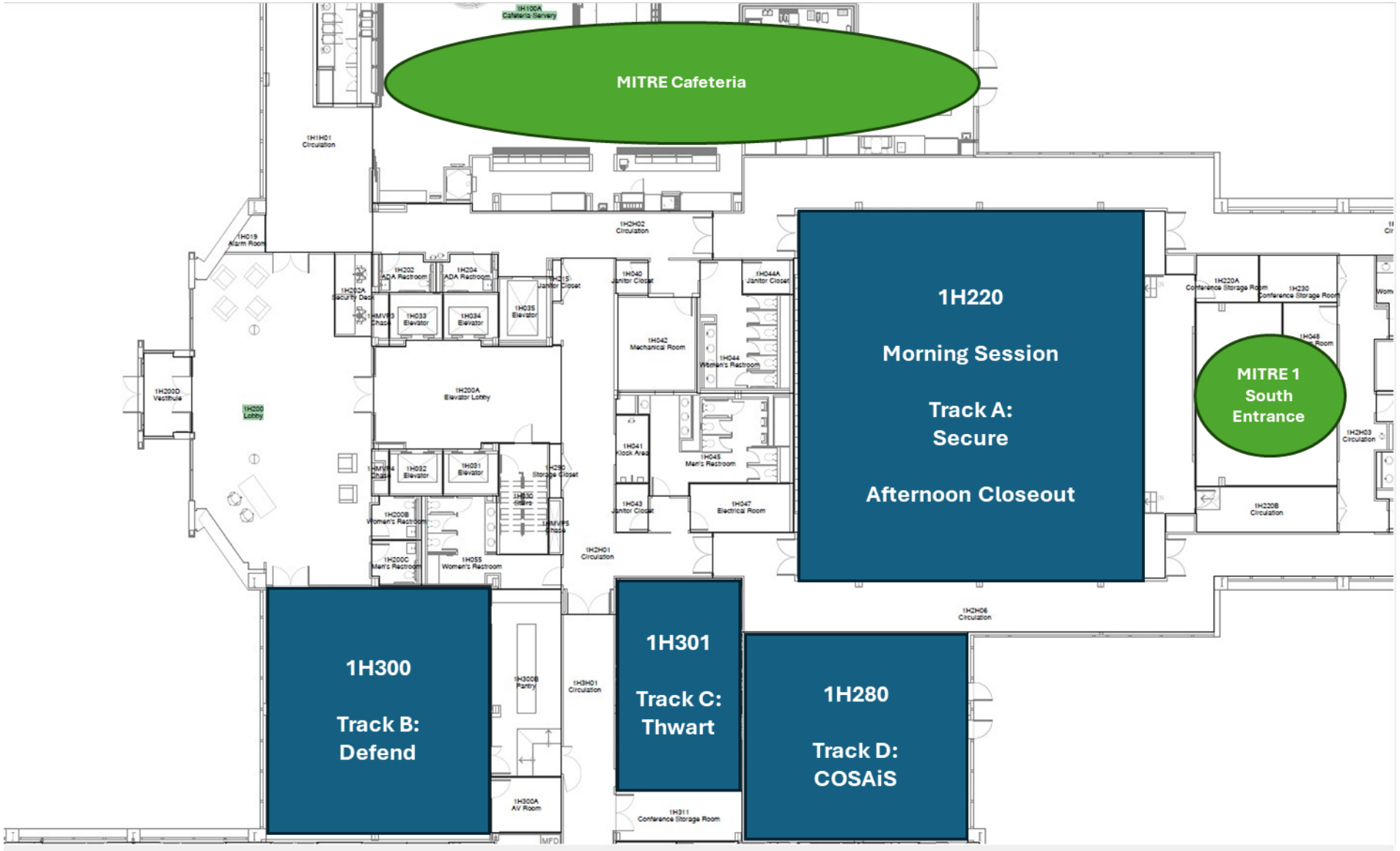
This Afternoon's Agenda

Time	Topic
12:00 - 1:00	Lunch (On Your Own)
1:00 - 2:05	Breakout Sessions – Round #1
2:05 - 2:15	Transition/Break
2:15 - 3:20	Breakout Sessions – Round #2
3:20 - 3:30	Transition/Break
3:30 - 4:35	Breakout Sessions – Round #3
4:35 – 4:40	Transition
4:40 - 5:00	Close-out

Breakout Session Topics:

- Track A: Secure Focus Area (1H220)
- Track B: Defend Focus Area (1H300)
- Track C: Thwart Focus Area (1H301)
- Track D: COSAiS (1H280)

Breakout Session Map



How You Contribute Today

- **Please raise your hand contribute**
- Members of the press, please identify yourself and your organization
- Be respectful of others
- Please don’t be shy – we would love to hear from everyone!
- **Please silence phones**
- **Recording is prohibited**

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

THANK YOU

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

CyberAIProfile@nist.gov

nccoe.nist.gov

@NISTcyber

Track A – Secure

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

How You Contribute Today

- **Please raise your hand contribute**
- Members of the press, please identify yourself and your organization
- Be respectful of others
- Please don’t be shy – we would love to hear from everyone!
- **Please silence phones**
- **Recording is prohibited**

Securing AI System Components (Secure)

Description: The Secure Focus Area supplements existing cybersecurity and risk management best practices to address novel, expanded, and altered attack surfaces associated with the integration of AI systems into an organization, its ecosystems, and its infrastructure. This covers the AI systems themselves, their supply chains, including data and machine learning infrastructure, and the other systems and data that the AI system relies on.

Themes from comments (so far):

- Add additional guidance for evidence-based, auditable AI tracking, testing, and sandboxing.
- Add guidance on handling multi-agent, multi-layered, and dynamic autonomous systems.
- Cover expanded threat surface caused by tool use and associated mitigations.

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

Examples of Challenges to Secure

- How do we handle the non-deterministic nature of AI?
- Who is accountable for the actions and decisions of AI?
- What actions and decisions should AI be allowed to make autonomously
- How do we keep private (training) data from being leaked?
- Data integrity (poisoning attacks, data origin, data quality)

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

Secure Discussion Questions

- How are organizations handling accountability for AI-driven decisions? What is working well and what challenges still exist?
- How can the Profile better capture and support the dynamic and rapidly-changing nature of AI?
- Are there additional considerations for more complex systems that deserve more attention (tool usage, multi-agent, embodied systems, etc.)?
- Where is more specificity / detail needed?

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

THANK YOU

We appreciate your participation in the Secure Breakout Session!

Visit the Cyber AI Profile project page for information on how to submit comments.

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Track B – Defend

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

How You Contribute Today

- **Please raise your hand contribute**
- Be respectful of others
- Please don’t be shy – we would love to hear from everyone!
- **Please silence phones**
- **Recording is prohibited**
- Members of the press, please identify yourself and your organization

Conducting AI-enabled Cyber Defense (Defend)

Description: The Defend Focus Area concentrates on identifying opportunities for the use of AI in supporting cybersecurity processes and activities and understanding the challenges that come with leveraging AI to assist in defensive operations. Opportunities to enhancing cyber defense capabilities using AI include mission assurance, predictive and proactive applications, investigation and analysis, and response and remediation.

Themes from comments (so far):

- Protection and use of metadata for effective AI-enabled defense
- Opportunities to use AI to increase rapid-response protections
- Relationship between AI-assisted or AI-enhanced applications and traditional mechanisms in the context of their purpose and objectives

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

Examples of Challenges to Defend

- Lacking explainability or provision of evidence, reasoning, or sources
- Producing confidently stated but erroneous or false content (“hallucinations” or “fabrications”)
- Exposing sensitive data in developing or training AI defense capabilities or models
- Protecting against information integrity and quality concerns due to training deviations, model drift or degradation, sabotage or manipulation, attacks, or lack of vendor transparency
- Providing ability to audit integrations and ensure system behavior
- Establishing responsibility and accountability governance for AI recommendations or decisions
- Integrating, standardizing, and training on human-AI teaming (e.g., security testing [red teaming] and traditional security testing]
- Balancing between human oversight and automation (e.g., Human-in-the-loop vs. Human-on-the-loop)
- Handling resource capacity requirements

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

Defend Discussion Questions

- What additional opportunities for using AI to improve cybersecurity capabilities should the Profile discuss?
- What considerations (e.g., hallucination risks, trust bias) must be reflected in the use of AI for enhancing cybersecurity capabilities to support effective cybersecurity defense?
- Where is more specificity / detail needed?

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

THANK YOU

We appreciate your participation in the Defend Breakout Session!

Visit the Cyber AI Profile project page for information on how to submit comments.

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Track C – Thwart

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

How You Contribute Today

- **Please raise your hand contribute**
- Members of the press, please identify yourself and your organization
- Be respectful of others
- Please don’t be shy – we would love to hear from everyone!
- **Please silence phones**
- **Recording is prohibited**

Thwarting AI-Enabled Cyber Attacks (Thwart)

Description: The Thwart Focus Area addresses how AI can enhance adversary capabilities, how these attacks could impact the entire cybersecurity landscape, and what organizations can do to bolster their systems against these emerging threats.

Themes from comments (so far):

- Address deepfake and synthetic media threats directly in the CSF Subcategories
- Incorporate/address rapid exploitation response
- Incorporate/address continuous improvement and audit-ready governance

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

Examples of Challenges to Thwart

AI-enabled attacks have the potential to scale and automate:

1. **Phishing and Social Engineering:**
   - **DeepFakes (vishing)**
2. **Vulnerability Discovery and Exploitation:**
3. **Supply Chain Attacks**
4. **Improved and Adaptive Malware**

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

Thwart Discussion Questions

How are you seeing adversaries use AI?

- AI Risk Database
- Anthropic Report
- CrowdStrike 2025 Threat Report

What aspects of Cybersecurity / Defense are at most risk to AI-powered attacks?

- Vulnerability management?
- Endpoint Detection?
- Email Security and Verification?

Where is more specificity / detail needed?

Discussion Topics:

- Clarify description
- Identify additional considerations
- Examine proposed priorities
- Capture any gaps
- Identify Informative References and other resources
- Resources for adoption and use

THANK YOU

We appreciate your participation in the Thwart Breakout Session!

Visit the Cyber AI Profile project page for information on how to submit comments.

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Track D – COSAiS (AI Overlays)

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

How You Contribute Today

- **Please raise your hand contribute**
- Members of the press, please identify yourself and your organization
- Be respectful of others
- Please don’t be shy – we would love to hear from everyone!
- **Please silence phones**
- **Recording is prohibited**

COSAiS Discussion Questions

Foundational Assumption: Using the NIST SP 800-53B Moderate Baseline

Objective: Test the core assumption and explore alternative risk-anchoring strategies

1. What implicit assumptions are we making by anchoring the overlay to the SP 800-53B Moderate impact baseline?
2. From your experience, does the Moderate baseline adequately capture the risk profile of AI-enabled systems? Why or why not?
3. What are the strengths of using an established baseline? Where does it potentially constrain innovation or miss emergent AI risks?
4. What alternative approaches could NIST consider?

Discussion Topics:

- **SP 800-53B moderate impact baseline**
- Updates to use cases and value add
- Overlay structure and control selection
- Deep Dive: worked examples
- Parking Lot Controls
- Gaps and Path Forward

COSAiS Discussion Questions

Use Cases: What's the Value Add?

Objective: Explore how use cases can add value to the overlays without becoming overly prescriptive.

1. Feedback on the consolidation of 4 use cases into 2.
2. Given that NIST guidance typically avoids deeply prescriptive use cases, what role should "real-world AI examples" play in an overlay?
3. Beyond grounding the reader, how might predictive AI use cases help:
   a. Clarify control intent?
   b. Demonstrate scope boundaries?
   c. Illustrate risk tradeoffs?

Discussion Topics:

- SP 800-53B moderate impact baseline
- **Updates to use cases and value add**
- Overlay structure and control selection
- Deep Dive: worked examples
- Parking Lot Controls
- Gaps and Path Forward

COSAiS Discussion Questions

Overlay Structure and Control Selection

Objective: Validate clarity, usability, and appropriateness of the overlay's structure.

1. Does the current structure help you quickly understand what is AI-specific versus inherited from existing controls? If not, what suggestions do you have for the structure?
2. Is the level of detail in the overlay appropriate for its intended audience (policy makers vs. implementers vs. assessors)?
3. What information is most valuable to you when consuming an overlay:
   a. Rationale for inclusion
   b. Tailoring guidance
   c. Implementation examples
   d. Link to potential attacks and mitigations (NIST AI 100-2e2025)
4. Are there areas where the overlay feels too abstract—or conversely, too detailed?

Discussion Topics:

- SP 800-53B moderate impact baseline
- Updates to use cases and value add
- **Overlay structure and control selection**
- Deep Dive: worked examples
- Parking Lot Controls
- Gaps and Path Forward

COSAiS Discussion Questions

Deep Dive: Worked Control Examples

Objective: Gather concrete, practitioner-level feedback.

1. What aspects of these worked examples are most helpful in understanding how AI changes control implementation?
2. How well do these examples clearly articulate what is “different” or “additional” for AI-enabled systems?
3. What is missing for implementers:
    1. Model lifecycle considerations?
    2. Data provenance and drift?
    3. Human oversight and accountability?
4. If you were system owner or implementing organization, would these examples help you make defensible risk decisions? Why or why not?

Discussion Topics:

- SP 800-53B moderate impact baseline
- Updates to use cases and value add
- Overlay structure and control selection
- **Deep Dive: worked examples**
- Parking Lot Controls
- Gaps and Path Forward

COSAiS Discussion Questions

Deep Dive: Worked Control Examples

Objective: Gather concrete, practitioner-level feedback.

Configuration Management

Control ID: *CM-02, Baseline Configuration*
Selected in SP 800-53B Moderate Baseline: *Yes*
Applicable AI Lifecycle Phase(s): *Model Training; Model Maintenance*
Assumptions: AI system configuration management may be integrated into organization-wide (enterprise).
Control Tailoring: [Discussion] Predictive AI systems may introduce additional configuration elements that are not explicitly addressed in baseline configurations of enterprise IT systems to potentially include machine learning frameworks and libraries, model architectures, and certain compute environments. The organization considers AI components and dynamic models in scope for the baseline configuration, including infrastructure components, the AI software stack, data and pipeline configurations, and model configuration artifacts.
Relevant NIST AI 100-2e2025 Attack ID: *NISTAML.024, NISTAML.026*

Discussion Topics:

- SP 800-53B moderate impact baseline
- Updates to use cases and value add
- Overlay structure and control selection
- **Deep Dive: worked examples**
- Parking Lot Controls
- Gaps and Path Forward

COSAiS Discussion Questions

Deep Dive: Worked Control Examples

Objective: Gather concrete, practitioner-level feedback.

System and Services Acquisition

Control ID: *SA-11(02), Developer Testing and Evaluation | Threat Modeling and Vulnerability Analyses*
Selected in SP 800-53B Moderate Baseline: *No*
Applicable AI Lifecycle Phase(s): *Model Training; Model Deployment*
Assumptions: The threat modeling and vulnerability analysis methods applied to the AI system are consistent whether the development and training of the AI system is performed within the system authorization boundary, performed by a separate development group within the organization, or is performed as part of a product delivered by a third-party vendor.
Control Tailoring: *[Discussion]* From development to deployment, developers update their understanding of the threats that can affect the deployment and track the vulnerabilities associated with the model, data sets, and tools used to create the predictive AI system environment. Given the scope of Predictive AI systems, additional focus may be applied to the system services supporting the machine learning.
Relevant NIST AI 100-2e2025 Attack ID: *NISTAML.013, NISTAML.022, NISTAML.05, NISTAML.051*

Discussion Topics:

- SP 800-53B moderate impact baseline
- Updates to use cases and value add
- Overlay structure and control selection
- **Deep Dive: worked examples**
- Parking Lot Controls
- Gaps and Path Forward

COSAiS Discussion Questions

Deep Dive: Worked Control Examples

Objective: Gather concrete, practitioner-level feedback.

System and Communications Protection

Control ID: *SC-07(10), Boundary Protection | Prevent Exfiltration*
Selected in SP 800-53B Moderate Baseline: *No*
Applicable AI Lifecycle Phase(s): *Model Deployment; Model Maintenance*
Assumptions: The impact from data exfiltration may be lower during development and training activities where exfiltration may be preventable using IT methods. In the deployment phase, the AI system boundary protections are implemented to prevent the exfiltration of the deployed model algorithms, data structures, and training data as a function of application usage.
Control Tailoring: *[Discussion]* The exfiltration of information about the training model can lead to the development of more sophisticated attacks that subvert the ability of the model to meet objectives and requirements. Additional protections are needed to prevent the exfiltration of information that can be used to engineer data privacy attacks and model privacy attacks. Tests for exfiltration can include methods to identify potential seeding by an attacker, identifying where an attacker is able to introduce data into the model that enables the extraction of additional information from the training model data sets. The frequency of testing may depend on the sensitivity of the data being used in the model and the criticality of the system relative to organization and system objectives for the system. *"extraction attacks are more successful when the model is seeded with more specific and complete information — the more the attacker knows, the more they can extract"* (NIST AI 100-2e2025, 3.3.2)
Relevant NIST AI 100-2e2025 Attack ID: *NISTAML.03, NISTAML.031, NISTAML.032, NISTAML.033, NISTAML.034)*

Discussion Topics:

- SP 800-53B moderate impact baseline
- Updates to use cases and value add
- Overlay structure and control selection
- **Deep Dive: worked examples**
- Parking Lot Controls
- Gaps and Path Forward

COSAiS Discussion Questions

Parking Lot Controls: Scope and Scale

Objective: Get feedback and first impressions on the scope and scale of the controls proposed for inclusion in the overlay.

1. Looking at the parking lot list, are the controls:
   a. Overinclusive?
   b. Underinclusive?
   c. Appropriate?
   d. TBD based on tailoring?
2. there a “magic number” of controls where the overlay remains usable without becoming overwhelming?
3. How should NIST prioritize controls for inclusion?

Discussion Topics:

- SP 800-53B moderate impact baseline
- Updates to use cases and value add
- Overlay structure and control selection
- Deep Dive: worked examples
- **Parking Lot Controls**
- Gaps and Path Forward

COSAiS Discussion Questions

Gaps and Path Forward

1. Are there AI-related cybersecurity or privacy risks you encounter that are not well addressed by SP 800-53 today (i.e., are there any gaps in the SP 800-53 controls that are not technology-specific)?
2. Should these risks be addressed through:
   a. The AI overlay only?
   b. Updates to the SP 800-53 catalog itself?
   c. Companion resources?
3. How can NIST balance stability of the catalog with the rapid evolution of AI risk?

Discussion Topics:

- SP 800-53B moderate impact baseline
- Updates to use cases and value add
- Overlay structure and control selection
- Deep Dive: worked examples
- Parking Lot Controls
- **Gaps and Path Forward**

THANK YOU

We appreciate your participation in the COSAiS Breakout Session!

View the draft annotate outline and share your comments.

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Workshop Closeout

Katerina Megas, NIST

Breakout Session Summaries

Track A	Track B	Track C	Track D
Secure	*Defend*	*Thwart*	*COSAiS*
Noah Schiro, MITRE	*Ishika Khemani, MITRE*	*Marissa Dotter, MITRE*	*Alicia Dawson, MITRE*

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Next Steps

- Analyze what we heard during this workshop as well as the public comments
- Identify any additional inputs needed to develop the initial public draft (IPD) of the Cyber AI Profile
- Start planning for potential COI meetings

Sign up for COI!
(scroll to bottom of page)

Submit Comments on the Preliminary Draft

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

THANK YOU

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

CyberAIProfile@nist.gov

nccoe.nist.gov

@NISTcyber