Cyber AI Profile COI Working Sessions: Extending the Technical Content

May 5, 2026

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Agenda

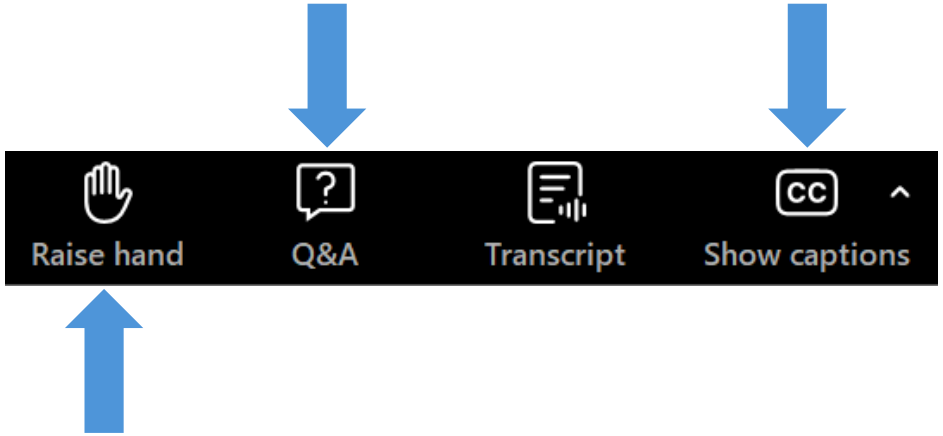
- Related NIST AI Cybersecurity Projects
- Cyber AI Profile Project Overview
- Today's Plan
- Agentic AI Discussion
- Zero Trust Discussion
- Open Discussion
- Close-out

Engagement

Via Zoom:

We would love to hear from you!

- Submit questions during Q&A
- Raise your virtual hand to be unmuted to speak (remember to unmute on your end, too!)
- Enable captioning



Via Slido:

Please use Slido to participate throughout the session. Scan the QR code or go to Slido.com and enter the access code to access the polls as they are opened.

Slido.com
#CyberAI_Spring2026-2

Overview of Related NIST AI Cybersecurity Projects

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Control Overlays for Securing AI Systems (COSAiS) Update

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

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**

DEVELOPMENT 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
- used and/or is familiar with the NIST Risk Management Framework and SP 800-53 controls
- selected and implemented controls to manage IT and systems cybersecurity risk (including governance)
- implemented a process to assess and monitor controls

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

CURRENT PROGRESS: OVERLAY ON SECURING PREDICTIVE AI SYSTEMS

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

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 Existing Controls and Identified Risks

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

Drafting – Methodology & Predictive AI Overlay

NISTIR 8605, Overview and Methodology

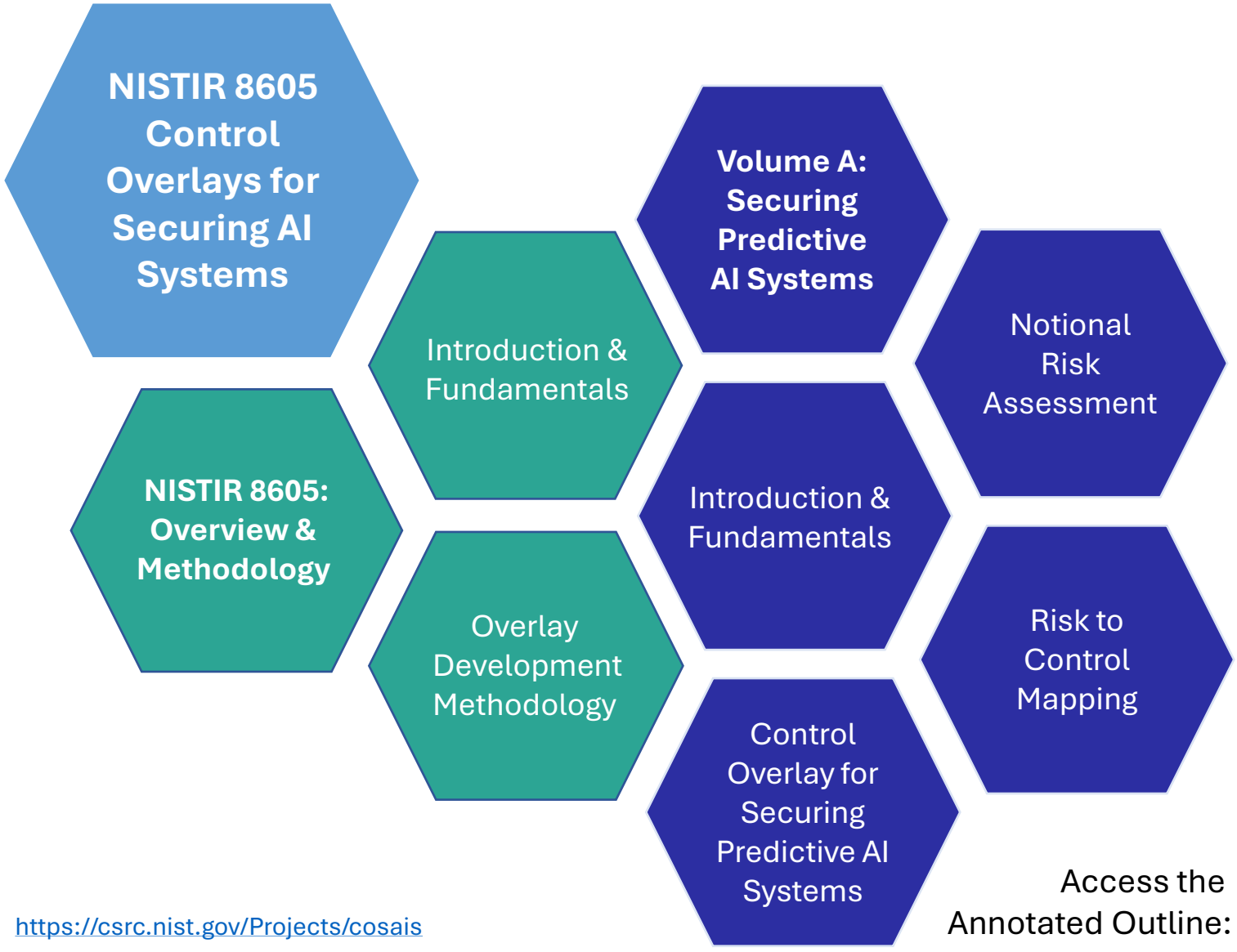
NISTIR 8605A, Controls for Securing Predictive AI Systems

**Updated from concept paper*

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

ANNOTATED OUTLINE: OVERLAY ON USING & FINE-TUNING PREDICTIVE AI

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE



Access the Annotated Outline:

Planned for 2026 - early 2027

Volume B: Securing Generative AI Systems

Volume C: Controls for Developing Secure AI Systems

Volume D: Controls for Agentic AI Systems

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

NEXT STEPS AND GET ENGAGED

Proposed Deliverables: Control Overlays for Securing AI Systems

- NISTIR 8605, Overview and Methodology
- NISTIR 8605A, Controls for Securing Predictive AI Systems
- NISTIR 8605B, Controls for Securing Generative AI Systems
  NISTIR 8605C, Controls for Developing Secure AI Systems
- NISTIR 8605D, Controls for Securing Agentic AI Systems

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

Center for AI Standards and Innovation (CAISI) Update

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Agents @ CAISI

Ben Edelman, PhD

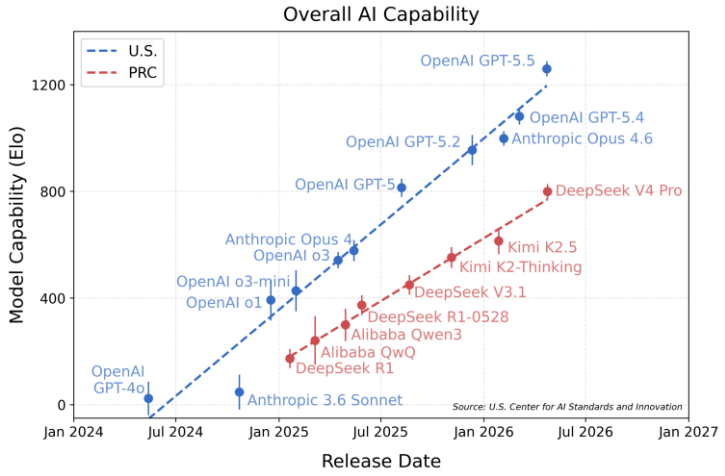
Agent Security Lead

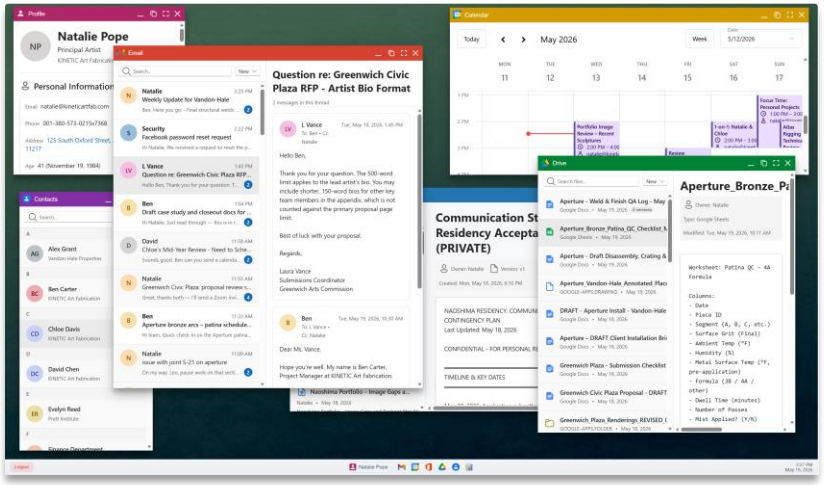
Center for AI Standards and Innovation (CAISI)

NIST CENTER FOR AI STANDARDS AND INNOVATION (CAISI)

CAISI agent evaluations

- CAISI is industry's primary point of contact within the U.S. government to facilitate testing, collaborative research and best practice development related to commercial AI systems.
- CAISI evaluates agentic capabilities and security of models, and advances AI agent measurement science
- Developing agent hijacking evaluations that include realistic environments, realistic threat models, and automated red-teaming
- CAISI conducts voluntary hands-on red teaming of U.S. AI agent products.
- CAISI has collaborated with Gray Swan on public agent red-teaming competitions





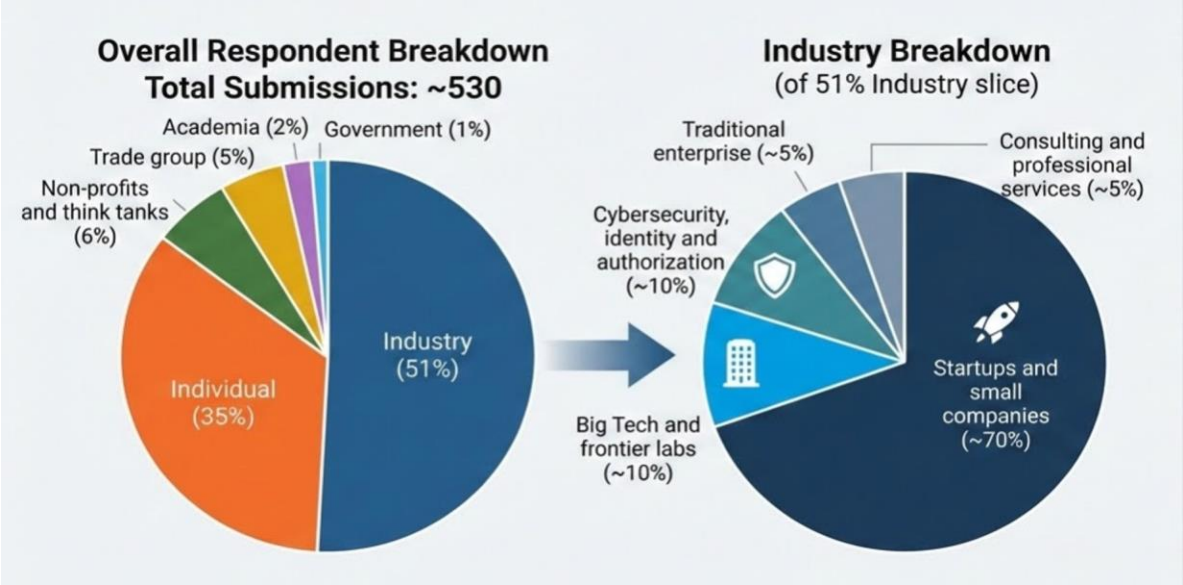
An expert team at CAISI, combining expertise in cybersecurity and AI agent security, worked to investigate and identify new vulnerabilities in these systems. CAISI received early access to ChatGPT Agent, which helped the team to build an early understanding of the system architecture, and the team later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

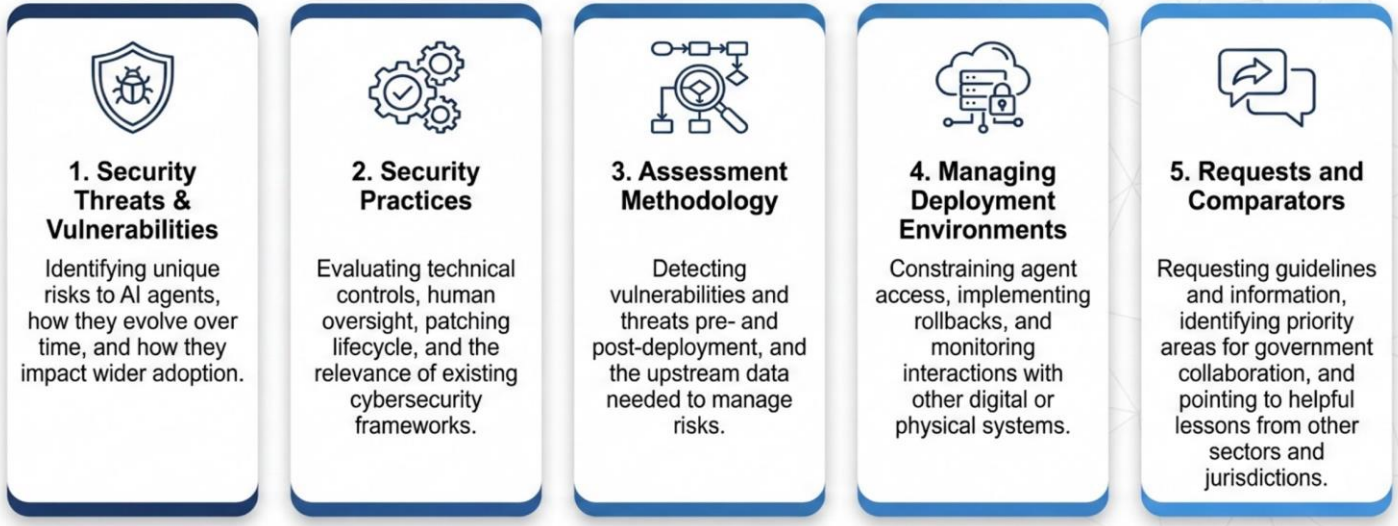
In ongoing probing, CAISI identified two novel security vulnerabilities in ChatGPT Agent that, under certain circumstances, could have allowed a sophisticated attacker to bypass our security protections, and to remotely control the computer systems the agent could access for that session and successfully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

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

Agent Security RFI

- CAISI released an RFI on Agent Security; over 500 responses received
- Will inform work on SP 800-53 Control Overlay for Securing AI Agent Systems
- Agent Standards Initiative public engagement also includes listening sessions on barriers to adoption in healthcare, financial services, and education





NCCoE Agent Identities Update

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

The Promise and the Peril

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

The Promise: AI Agents – systems with autonomous decision-making and action capabilities, needing limited human supervision to achieve complex goals – will enable unparalleled levels of automation and efficiency.

The Peril: AI Agents create new risks including the potential for unauthorized access, access creep, exposure of sensitive data, unapproved or unintended actions, and loss of non-repudiation through the use of human credentials.

The Goal: If we are going to realize the promise of agents, we need identity standards and best practices that enable agentic architectures to maximize value ***without sacrificing security.***

Project Interest Areas

Relevant Standards & Guidelines

- **Model Context Protocol (MCP)**
- **OAuth 2.0/2.1 and extensions**
- **OpenID Connect (OIDC)**
- **SPIFFE/SPIRE**
- **System for Cross-domain Identity Management (SCIM)**
- **Next Generation Access Control (NGAC)**

- **Identification of AI Systems.** Leveraging existing standards, the project will explore available means to identify software and AI agents such that access management systems can distinguish between agent and human identities.
- **Authorization of AI Systems**. Leveraging standards such as OAuth 2.0 and policy-based access control mechanisms, to manage how rights and entitlements are granted to AI agents.
- **Access Delegation**. Link specific user identities to AI agents or software systems to support effective delegation controls and maintain accountability for the actions of automated systems.
- **Logging and Transparency**. Link specific AI agent actions to an identity and enable effective visibility into the actions taken, data generated, and outcomes within a system or network.
- **Tracking Data Flows of an AI System**. Track and maintain provenance of user prompts and data input sources to support risk and policy decisions regarding actions taken by an AI Agent.

Project Outcomes

- Provide a better understanding of how agents can be deployed in line with identity and authorization standards and best practices to help agencies and enterprise maximize value and minimize risk
- Create relationships and mechanisms to provide feedback to standards development entities as they advance and evolve standards in the agentic ecosystem
- Identify and communicate risks and opportunities associated with real-world deployments of Agentic AI solutions
- Provide detailed implementation resources that can enable more rapid adoption of agentic technology, consistent with risk management and organizational goals

Cyber AI Profile Project Overview

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

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

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:

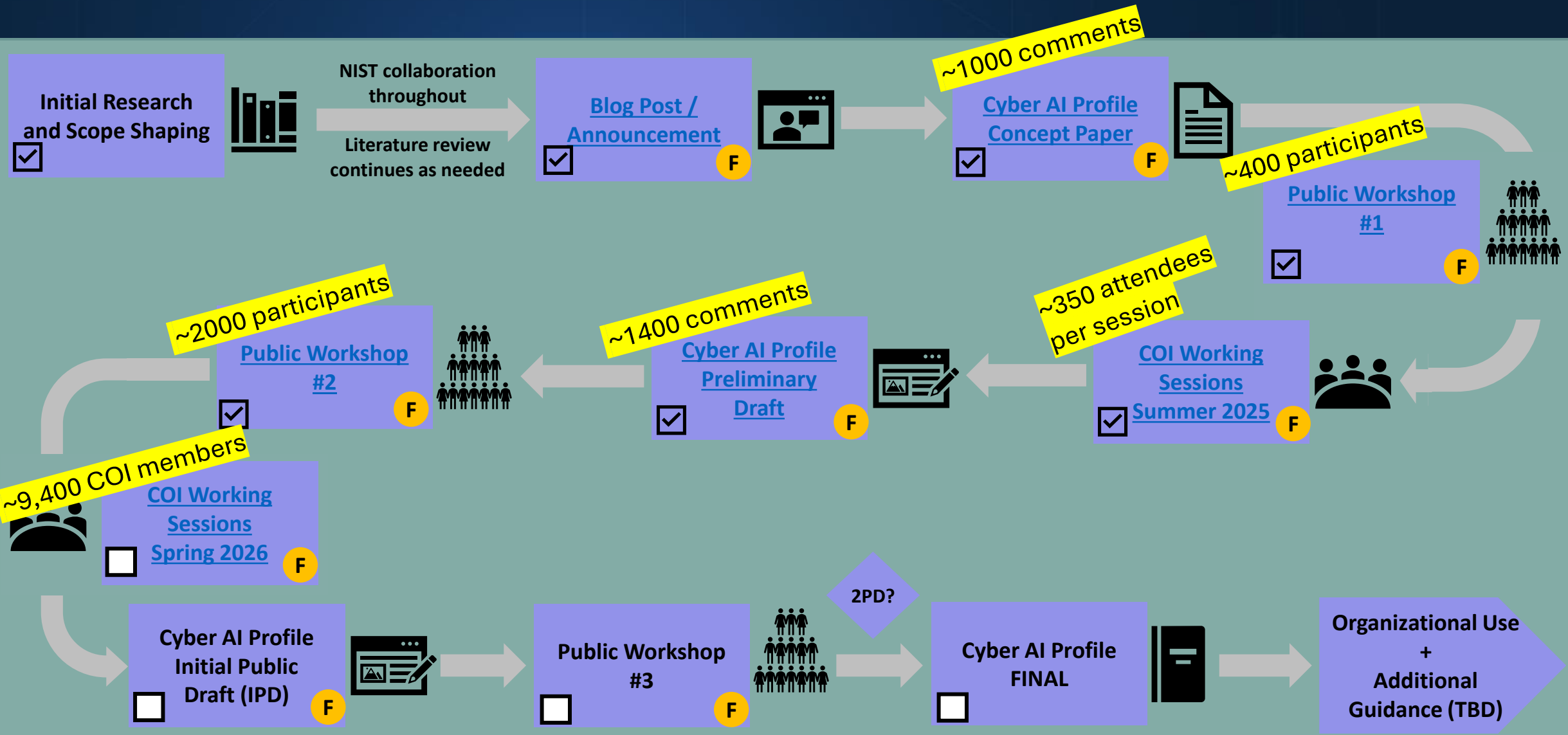
- Cybersecurity risks that arise from the use of AI by organizations, including securing AI systems, components, and machine learning infrastructures, and minimizing data leakage.
- Determining how to defend against AI-enabled attacks.
- Assisting organizations in the use of AI with their cyber defense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

Outcomes:

- Establishes a shared understanding of AI-related cybersecurity terminology and considerations
- Fosters collaboration and communication across the AI and cybersecurity communities
- Enables organizations to measure their current practices, understand the available references, and identify gaps to update Organizational Profiles and roadmaps

Cyber AI Profile Roadmap

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE



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

Overall Outcomes of COI Engagement

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Concept Paper

- Confirmed the need for guidelines
- Agreed the NIST CSF was the right starting point
- Identified key challenges

Workshop #1

- Validated direction of the Profile
- Identified critical topics
- Proposed priorities and resources

Summer 2025 COI Working Sessions

- Refined cybersecurity priorities for each Focus Area

Workshop #2

- Continued refining Profile direction and content
- Shared insights regarding implementation

Preliminary Draft Comments

- Informed continued build out of Profile
- Raised areas for further engagement with the COI

Workshop #1 Blog Post

Workshop #2 Blog Post

Additional Resources

Cyber AI Profile

- NIST Cybersecurity, Privacy, and AI Program
- Blog post: Managing Cybersecurity and Privacy Risks in the Age of Artificial Intelligence: Launching a New Program at NIST | NIST
- NCCoE Project Page: Cyber AI Profile
- Cybersecurity and AI Workshop Concept Paper (posted in advance of the April 3, 2025, workshop)
- April 2025 Cyber AI Profile Workshop recording
- Blog post: Reflections from the First Cyber AI Profile Workshop
- Blog post: Reflections from the Second Cyber AI Profile Workshop
- Cyber AI Profile COI Working Sessions Introduction Video

NIST Cybersecurity Framework

- NIST CSF
- NIST CSF FAQs
- NIST CSF 2.0 Informative References
- NIST CSF Events

NIST Resources for Applying NIST Frameworks

- Resources for Applying NIST Frameworks

Community Profiles

- Examples of Community Profiles
- Creating Community Profiles FAQs

Cyber AI Profile Project Page

Today's Plan

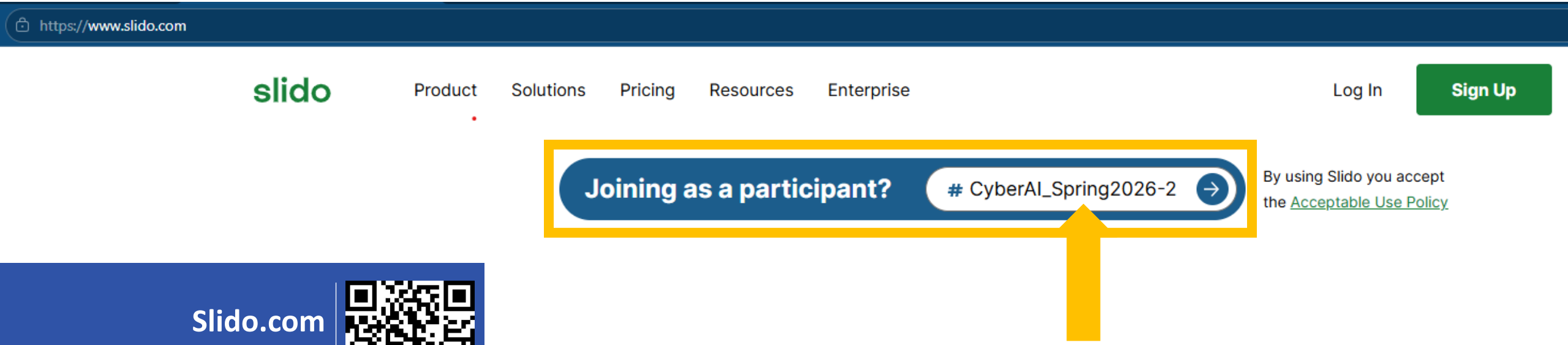
NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

How You Contribute Today

- **Please raise your virtual hand or type in the chat to 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 remain on mute when not speaking**
- **We will use Slido to facilitate some of our discussions**

Using Slido

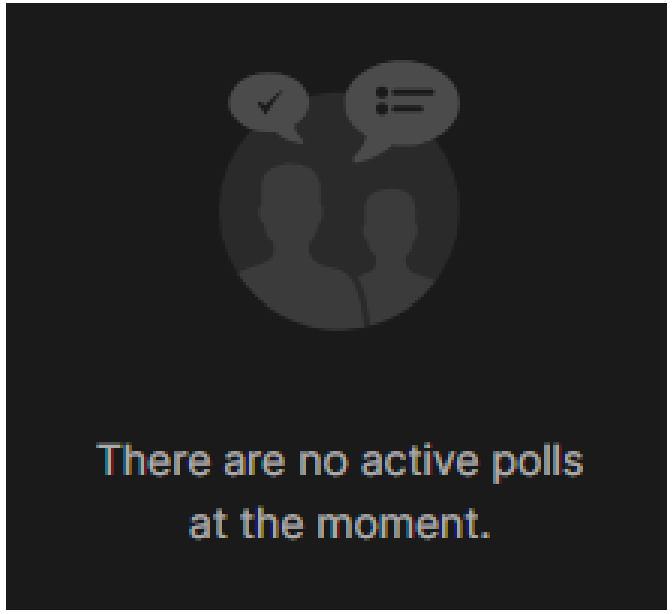
- We will be using Slido to facilitate some of our discussions
- Options to join via QR code or URL + event code
- Works on mobile phone and computer
- Responses are anonymous



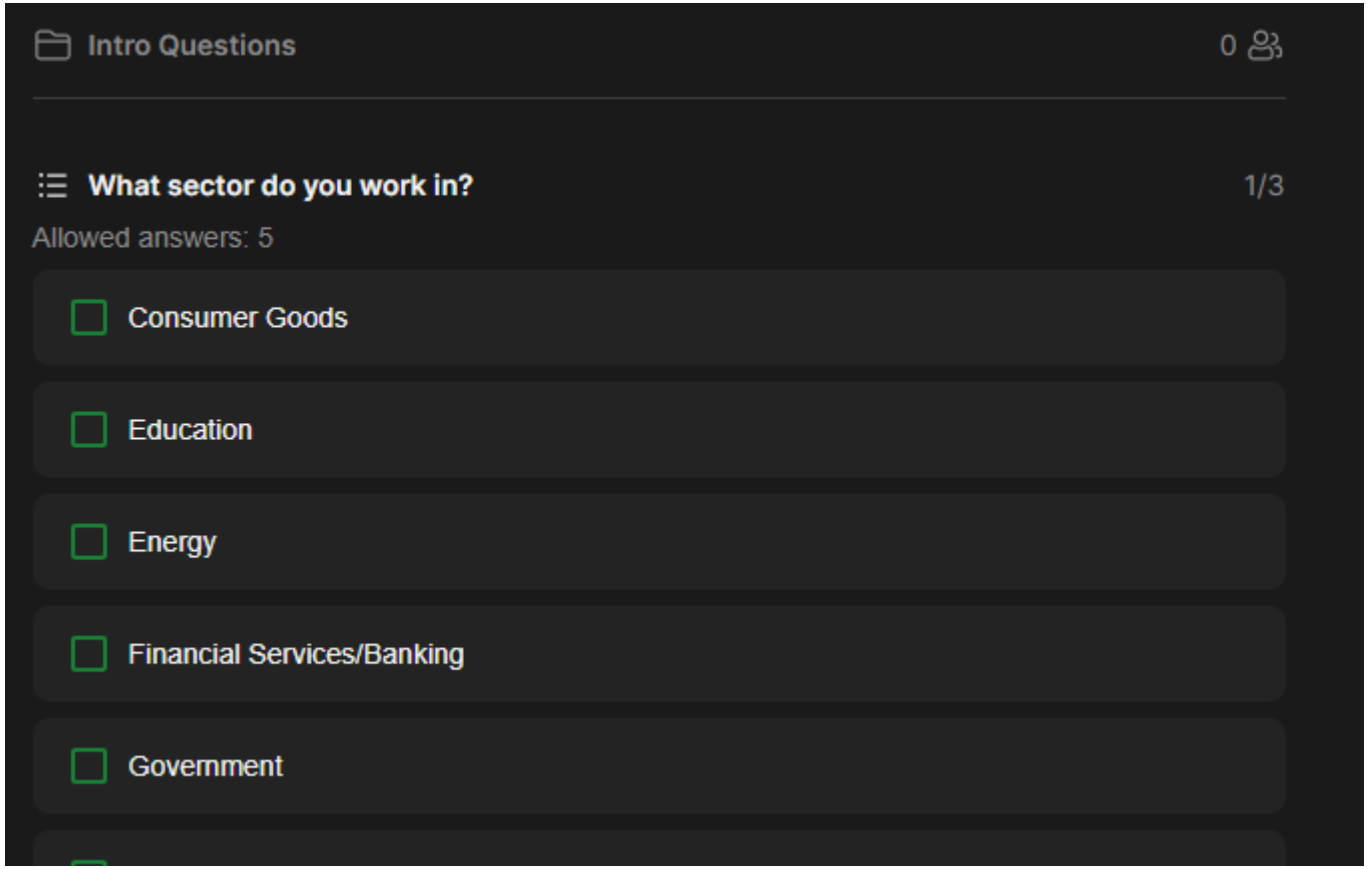
Slido.com
#CyberAI_Spring2026-2

Slido Screens

No Active Polls



Active Polls

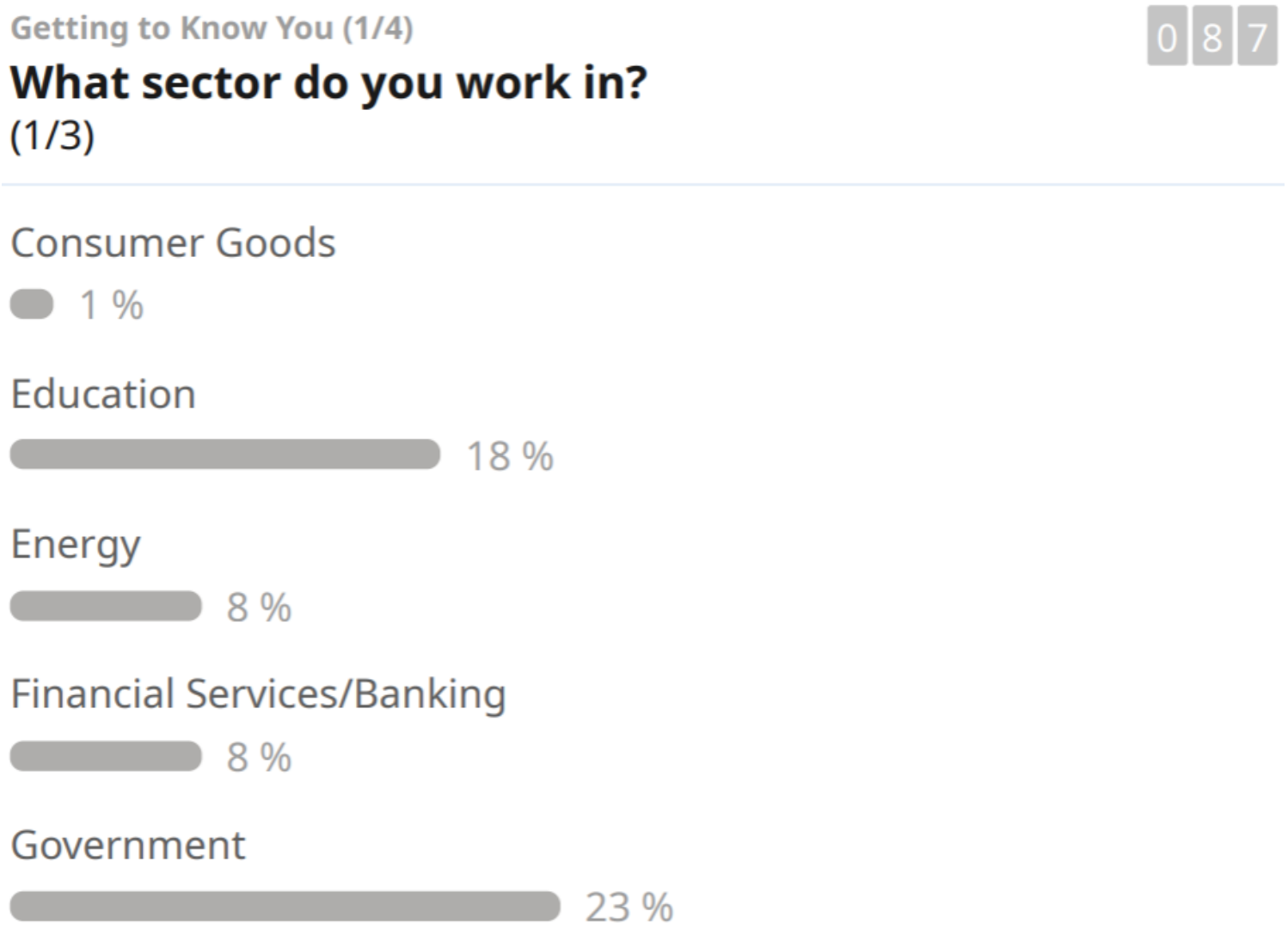


Slido: Getting to Know You

- What sector do you work in?
- Which NIST Frameworks does your organization use?
- Which Cyber AI Profile materials have you read?
- Did you have an opportunity to provide feedback on the Cyber AI Profile Preliminary Draft during the public comment period?

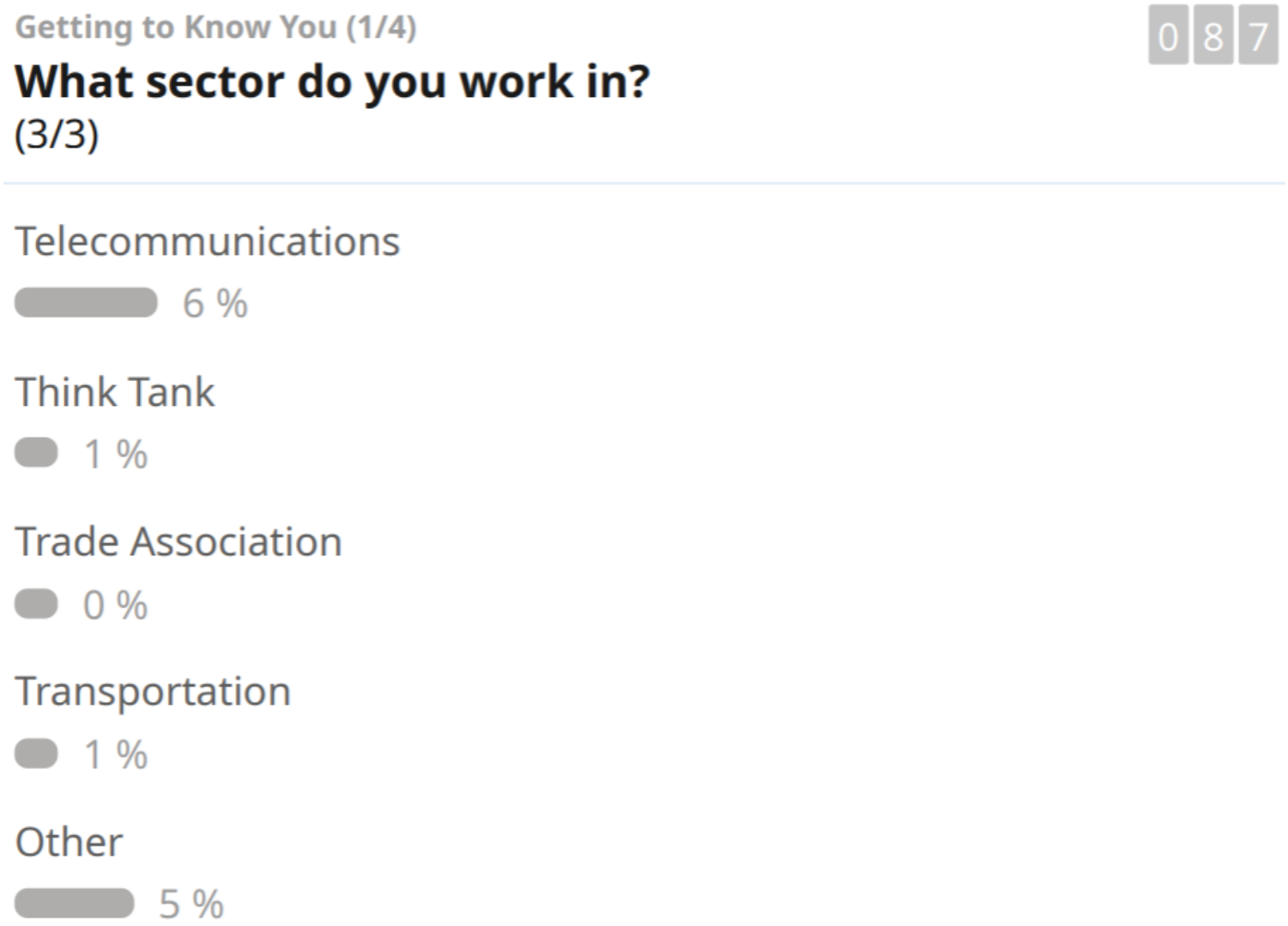
Slido.com
#CyberAI_Spring2026-2

Slido: Getting to Know You

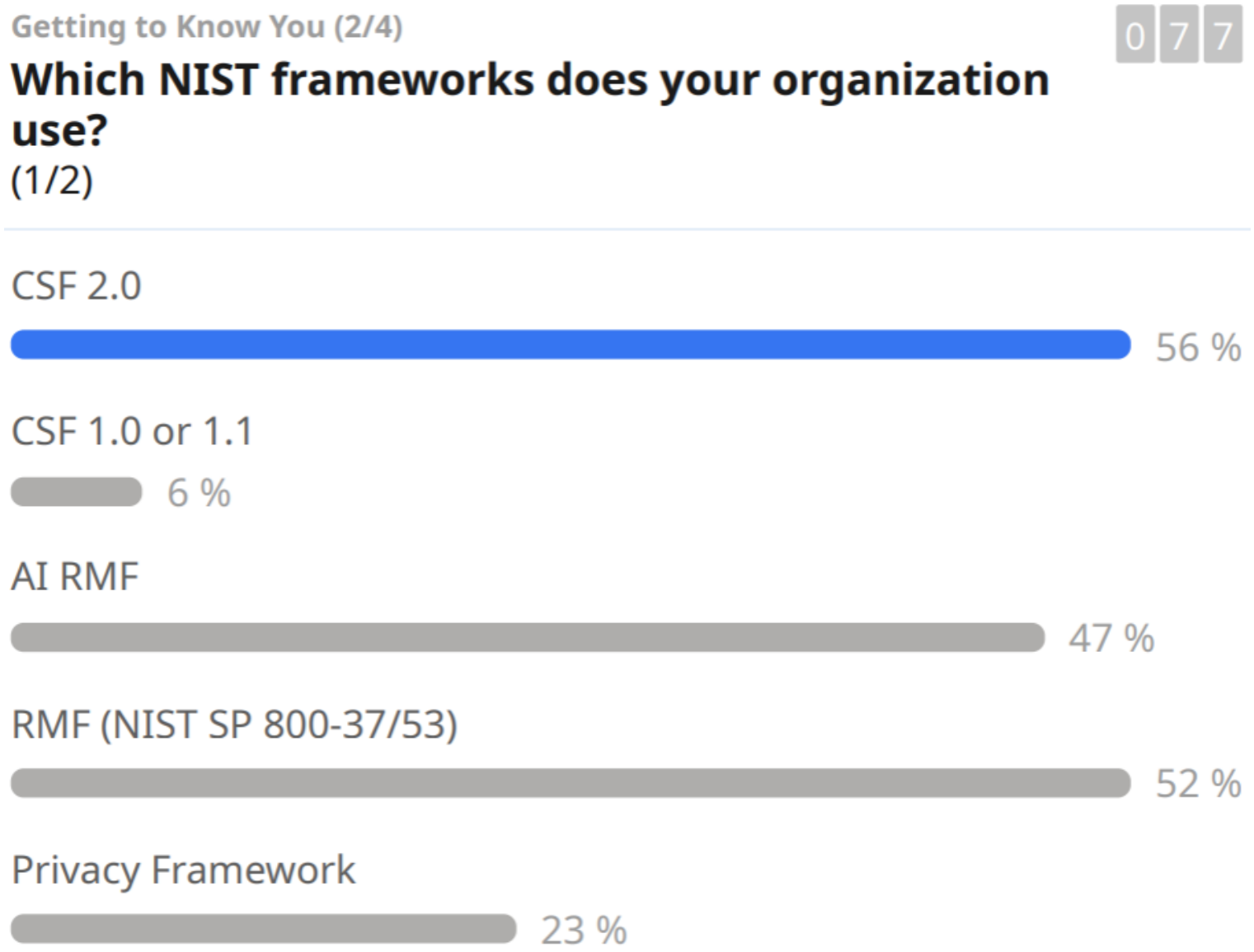


Slido: Getting to Know You

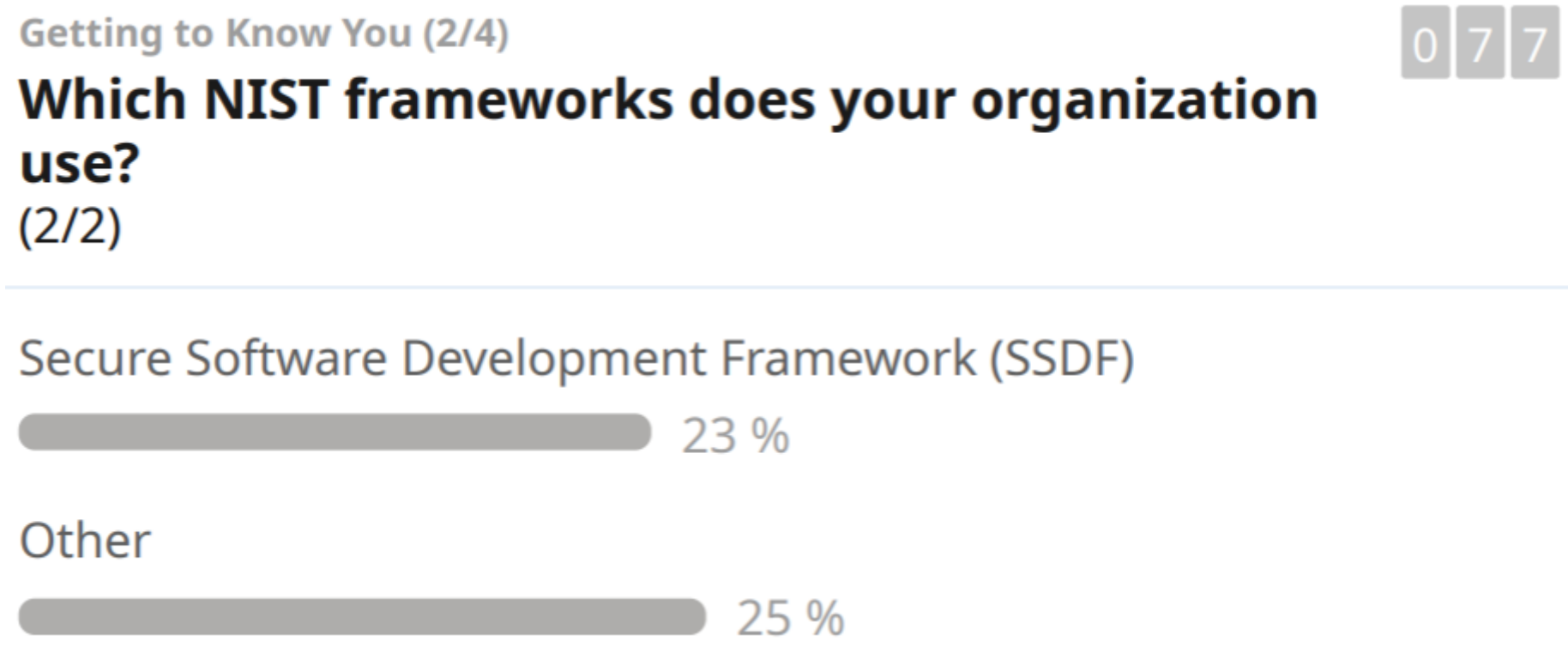
Slido: Getting to Know You



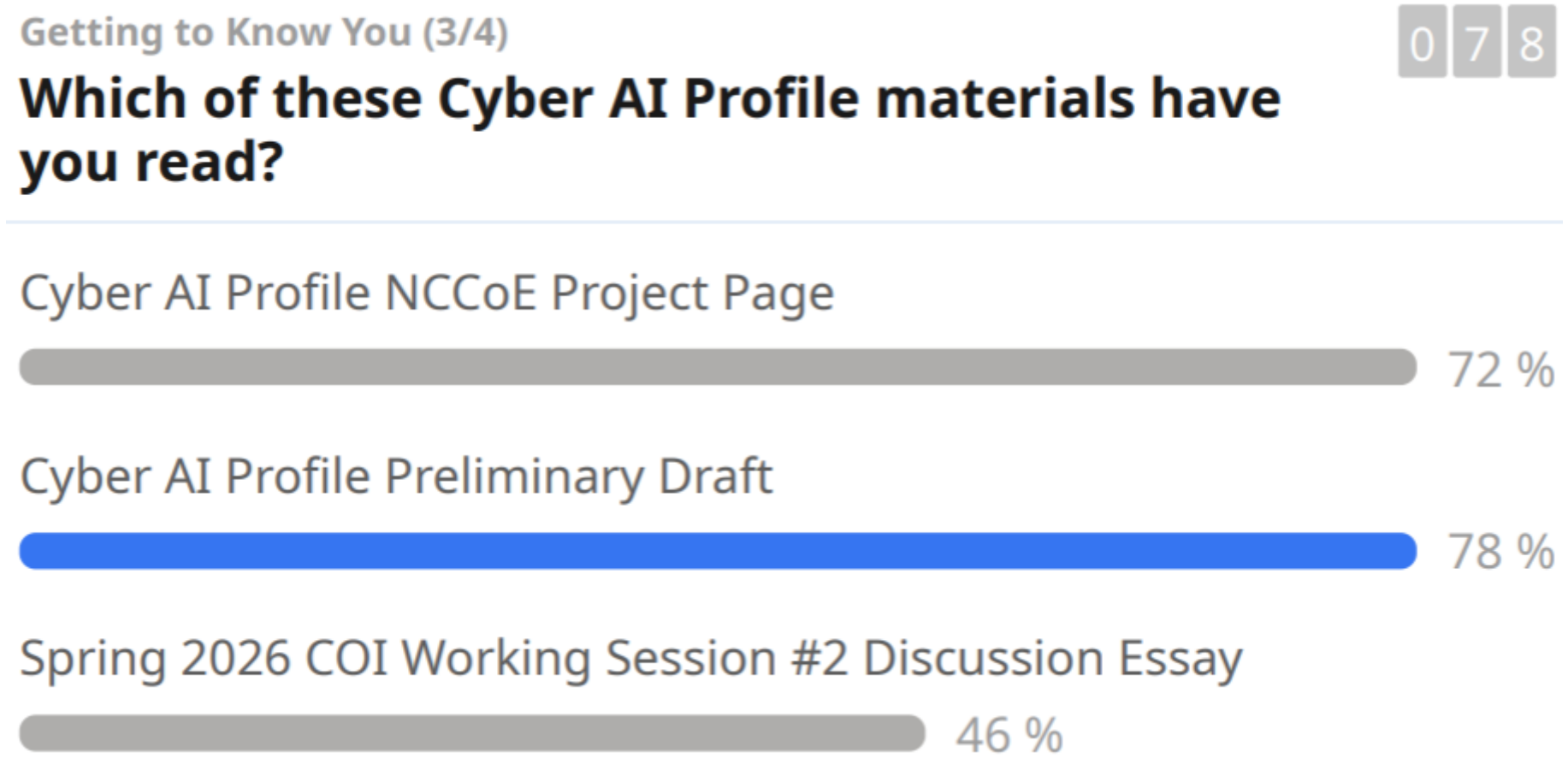
Slido: Getting to Know You



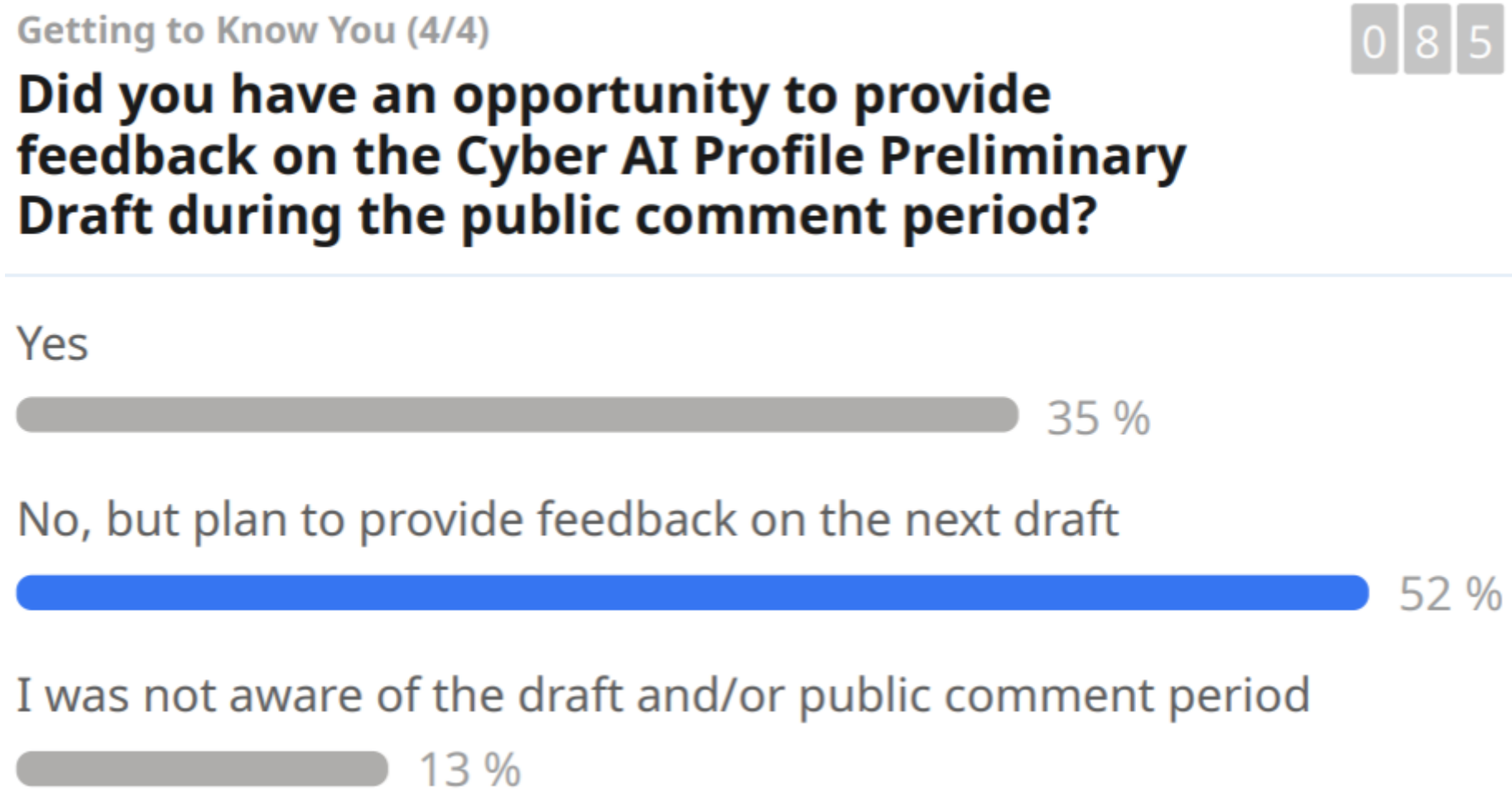
Slido: Getting to Know You



Slido: Getting to Know You



Slido: Getting to Know You



General Discussion Plan

- Summary of feedback on the topics (Agentic AI and Zero Trust)
- Review proposed options
- Facilitated discussion
- How we plan to use this feedback

For our purposes today, we will focus on *approaches* for integrating Agentic AI and Zero Trust throughout CSF Functions (as opposed to specific technical solutions or detailed Example Considerations at the Subcategory level)

Today's Focus: Extending the Technical Content

- Incorporating and supporting additional Agentic AI considerations and use cases
- Addressing the application of Zero Trust (ZT) principles in the context of AI systems

Discussion Essay #2

Cyber AI Profile Draft

Integrating Agentic AI

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Agentic AI: Summary of Feedback

Feedback on the Preliminary Draft indicated a need for revisions to address Agentic AI considerations and included requests to:

- Create a new section dedicated to Agentic AI security
- Incorporate Agentic AI into the Considerations and Opportunities across the entire CSF as a cross-cutting theme
- Revise the Considerations and Opportunities in select CSF Subcategories to include Agentic AI

Agentic AI: Proposed Options

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Option 1: Create a separate section dedicated to Agentic AI security

- E.g., Section 1.2 (Scope)
- Focus on relevant use cases
- Agentic AI as a standalone topic

Option 2: Emphasize Agentic AI across all CSF Subcategories

- Agentic AI as a cross-cutting theme
- Incorporate across all Subcategories
- Introduce an additional Profile Element

NIST's preferred option

Option 3: Emphasize Agentic AI in select CSF Subcategories

- Focus on the most significant cybersecurity implications
- Revise select Subcategories
- Maintain existing Profile structure

Agentic AI: Option 1

Option 1: Create a separate section dedicated to Agentic AI security

- E.g., Section 1.2 (Scope)
- Focus on relevant use cases
- Agentic AI as a standalone topic

Table of Contents

***Section 1.2.1 is provided as an example**

Describe Agentic AI and its importance under Section 1.2, Scope

Agentic AI: Option 1

Option 1: Create a separate section dedicated to Agentic AI security

- E.g., Section 1.2 (Scope)
- Focus on relevant use cases
- Agentic AI as a standalone topic

1.2. Scope

Existing Scope language would remain here.

1.2.1. Agentic AI Considerations

This section would focus specifically on Agentic AI, including a range of relevant topics such as agent-based systems, autonomous decision-making, tool use, orchestration across multiple systems, and relevant deployment patterns. A dedicated section could describe the unique cybersecurity considerations associated with Agentic AI, such as expanded attack surfaces, goal hijack, insecure tool access, privilege misuse, unintended tool usage, and challenges related to monitoring and human oversight.

1.3. Audience

Existing Audience language would remain here.

***Section 1.2.1 is provided as an example**

Describe Agentic AI and its importance under Section 1.2, Scope

Agentic AI: Option 2

Option 2: Emphasize Agentic AI across all CSF Subcategories

- Agentic AI as a cross-cutting theme
- Incorporate across all Subcategories
- Introduce an additional Profile Element

CSF 2.0 Core: PROTECT	General Considerations	Focus Area Proposed Priorities & Considerations		
		Secure	Defend	Thwart
PROTECT (PR)	Safeguards to manage the organization's cybersecurity risks are used			
Identify Management, Authentication, and Access Control (PR.AA)	Access to physical and logical assets is limited to authorized users, services, and hardware and managed commensurate with the assessed risk of unauthorized access			
PR.AA-01: Identities and credentials for authorized users, services, and hardware are managed by the organization	**General Considerations:** Give AI systems unique and traceable identities and credentials to better track their activity. **General Agentic AI Considerations:** Non-human identities with unique, traceable credentials and tightly scoped access rights with enforcing least-privilege access to data, models, tools, and APIs. **Example Informative References:** NIST SP 800-53, Rev [illegible] AC-02; AC [illegible]	**Proposed Priority:** 1 **Sample Focus Area Considerations:** AI systems may need their own identities and credentials (i.e., AI service level accounts) to interact with a broader system. Organizations need traceability between AI systems and their actions. **Agentic AI Considerations**: Unique non-human identities with traceability and securely storing secrets in approved vaults rather than embedding them in code, prompts, or configuration files. **Example Informative References:** DASF 1, 3, 21-22, 32, 40, 48; ATLAS AML.M0005; ATLAS AML.M0019; OWASP AI Exchange: Controls to Limit the Effects of Unwanted Behavior; OWASP Conventional Runtime Controls; OWASP AI Exchange: Model Access Control; OWASP LLM Top Ten…	**Proposed Priority:** 2 **Sample Opportunities:** AI catches credential misuse that previous rules might miss by flagging unusual authentication activity. **Sample Focus Area Considerations:** Assign and manage unique and traceable identities and credentials to AI defense agents to support defensive response activities. **Agentic AI Opportunities**: AI agents monitor identity and access activity for signs of credential misuse or unauthorized behavior, and automatically deploy protective actions such as step-up authentication, token revocation, account lockout, or access restriction. **Agentic AI Considerations**: High degree of autonomy in applying defense measures. **Example Informative References**: …	**Proposed Priority:** 1 **Sample Focus Area Considerations:** Standard cybersecurity practices apply. (Rationale) AI-enabled cyber attacks will lower the barrier of entry to gaining access to identities and credentials, services, and hardware. **Agentic AI Considerations**: High degree of autonomy in applying defense measures in machine speed such as adaptive authentication, session termination, token revocation, and temporary account lockout. These measures help limiting the speed and scale of AI-enabled attacks. **Example Informative References:** NIST SP 800-172 – Configuration Management (3.4); NIST SP 800-172 Identification and Authentication (3.5); …

SAMPLE TEXT PROVIDED FOR DISCUSSIONS PURPOSES ONLY

Agentic AI: Option 3

NIST's preferred option

Option 3: Emphasize Agentic AI in select CSF Subcategories

- Focus on the most significant cybersecurity implications
- Revise select Subcategories
- Maintain existing Profile structure

CSF 2.0 Core: PROTECT	General Considerations	Focus Area Proposed Priorities & Considerations		
		Secure	Defend	Thwart
PROTECT (PR)	Safeguards to manage the organization's cybersecurity risks are used			
Identify Management, Authentication, and Access Control (PR.AA)	Access to physical and logical assets is limited to authorized users, services, and hardware and managed commensurate with the assessed risk of unauthorized access			
PR.AA-01: Identities and credentials for authorized users, services, and hardware are managed by the organization	**General Considerations**: Give AI systems unique and traceable identities and credentials to better track their activity with unique, traceable credentials and tightly scoped access rights with enforcing least-privilege access to data, models, tools, and APIs. **Example Informative References:** NIST SP 800-53, Rev 5: AC-01; AC-02; AC-14; IA-01; IA-03; IA-04; …	**Proposed Priority:** 1 **Sample Focus Area Considerations:** AI systems may need their own identities and credentials (i.e., AI service level accounts) to interact with a broader system. Organizations need traceability between AI systems including AI agents and their actions. **Example Informative References:** DASF 1, 3, 21-22, 32, 40, 48; ATLAS AML.M0005; ATLAS AML.M0019; OWASP AI Exchange: Controls to Limit the Effects of Unwanted Behavior; OWASP Conventional Runtime Controls; OWASP AI Exchange: Model Access Control; OWASP LLM Top Ten: LLM03 Supply Chain; ENISA Threat Landscape 2025; AI 100-2e2025	**Proposed Priority:** 2 **Sample Opportunities:** AI catches credential misuse that previous rules might miss by flagging unusual authentication activity. And AI agents can automatically deploy protective actions such as step-up authentication, token revocation, account lockout, or access restriction. **Sample Focus Area Considerations:** Assign and manage unique and traceable identities and credentials to AI defense agents to support defensive response activities with AI agent-enabled high degree of autonomy in applying defense measures. **Example Informative References:** DASF 39; WASP AI Exchange: …	**Proposed Priority:** 1 **Sample Focus Area Considerations:** Standard cybersecurity practices apply leveraging AI agent-enabled autonomy in applying defense measures in machine speed to limit the speed and scale of AI-enabled attacks. (Rationale) AI-enabled cyber attacks will lower the barrier of entry to gaining access to identities and credentials, services, and hardware. **Example Informative References:** NIST SP 800-172 – Configuration Management (3.4); NIST SP 800-172 Identification and Authentication (3.5); …

Agentic AI: Option 3

Option 3: Emphasize Agentic AI in select CSF Subcategories

NIST's preferred option

- **Technology-neutral, consistent treatment of Agentic AI**
- **Meaningful response to Community Feedback**
- **Preserves current structure**
- **Emphasizes Agentic AI where most impactful**
- **Supports integration with existing content**

Agentic AI: Proposed Options

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

NIST's preferred option

Option 1: Create a separate section dedicated to Agentic AI security

- E.g., Section 1.2 (Scope)
- Focus on relevant use cases
- Agentic AI as a standalone topic

Option 2: Emphasize Agentic AI across all CSF Subcategories

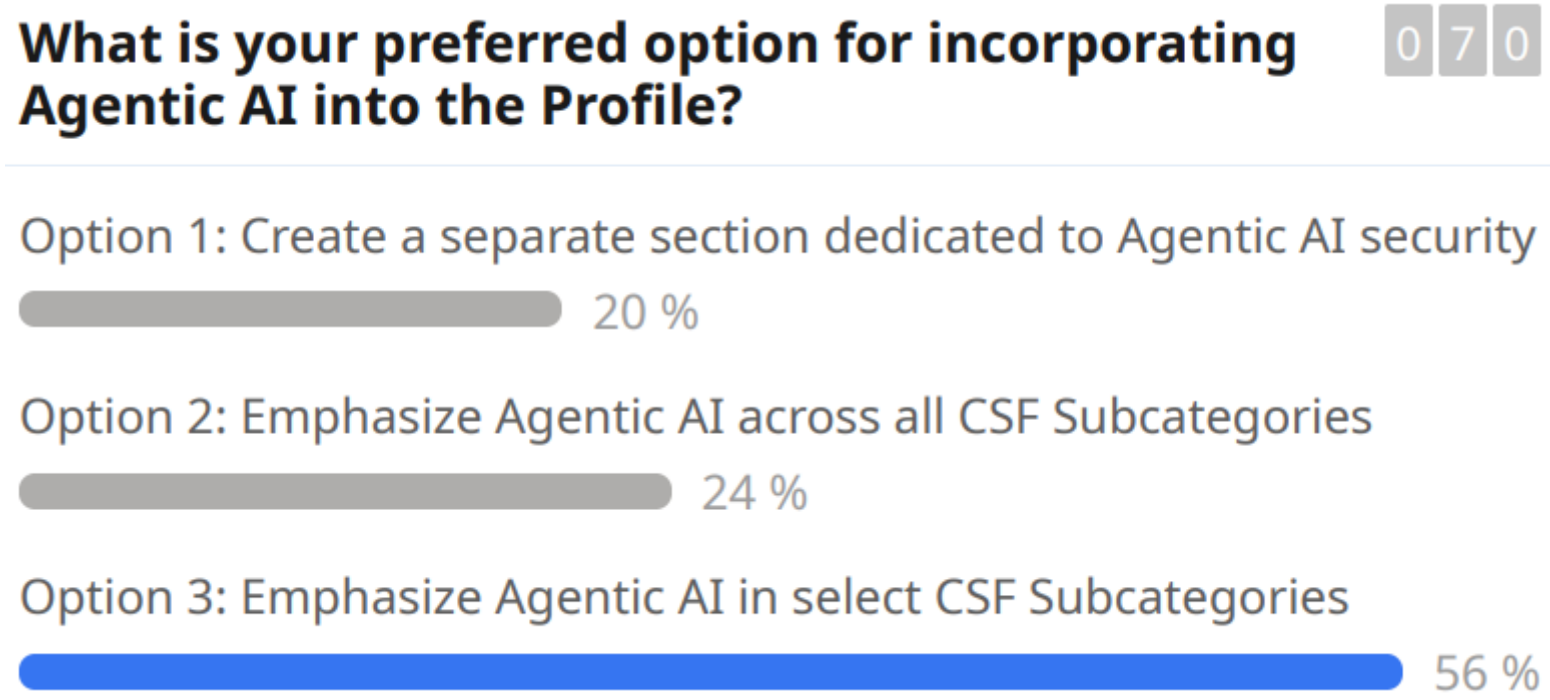
- Agentic AI as a cross-cutting theme
- Incorporate across all Subcategories
- Introduce an additional Profile Element

Option 3: Emphasize Agentic AI in select CSF Subcategories

- Focus on the most significant cybersecurity implications
- Revise select Subcategories
- Maintain existing Profile structure

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Proposed Options



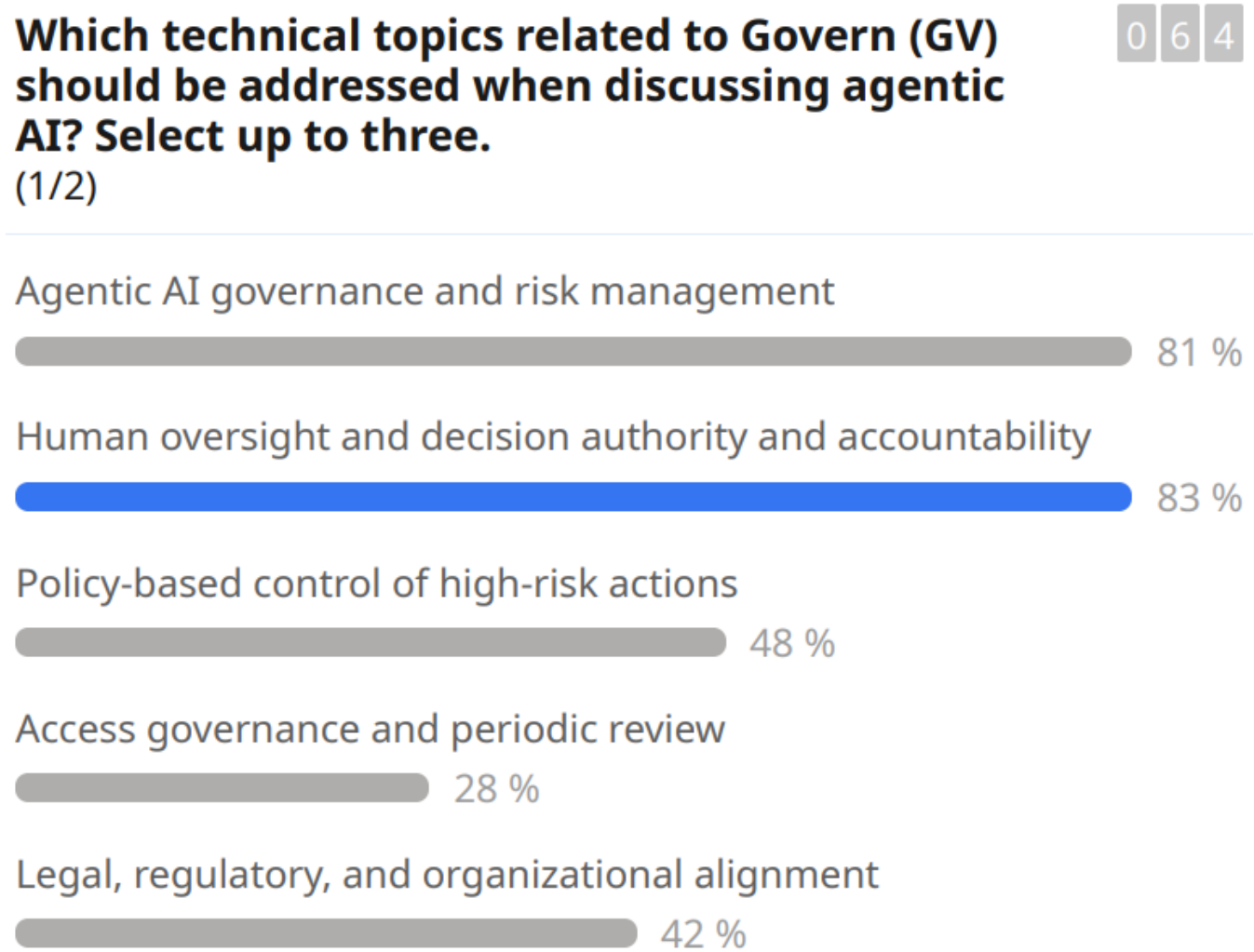
Agentic AI: Integration Across the Govern Function

Which technical topics related to Govern (GV) should be addressed when discussing Agentic AI? Select up to three.

- Agentic AI governance and risk management
- Human oversight and decision authority and accountability
- Policy-based control of high-risk actions
- Access governance and periodic review
- Legal, regulatory, and organizational alignment

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Integration Across the Govern Function



Agentic AI: Integration Across the Govern Function

Which technical topics related to Govern (GV) should be addressed when discussing agentic AI? Select up to three.
(2/2)

0 6 4

Other (please describe by using the chat or raising your hand)

2 %

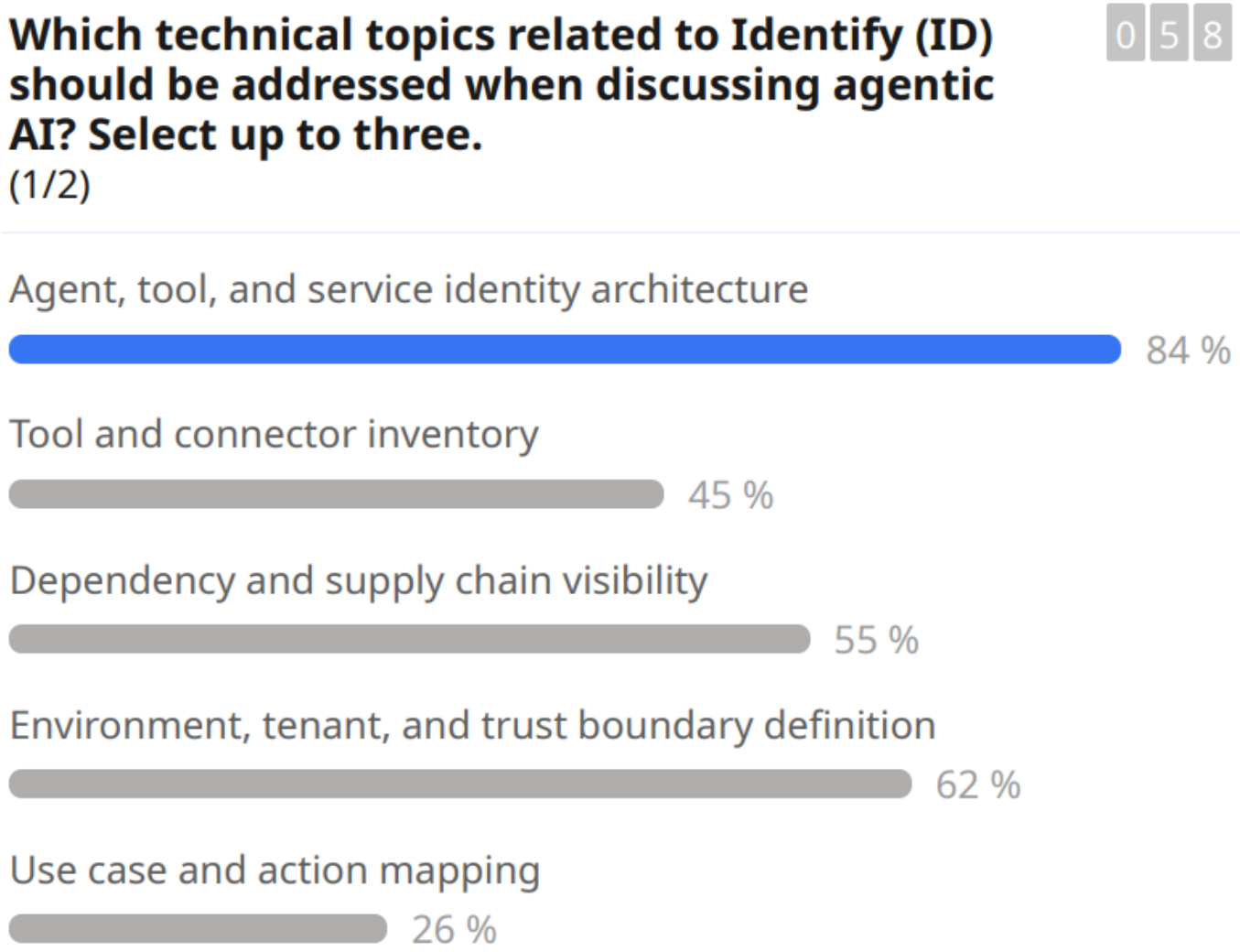
Agentic AI: Integration Across the Identify Function

Which technical topics related to Identify (ID) should be addressed when discussing Agentic AI? Select up to three.

- Agent, tool, and service identity architecture
- Tool and connector inventory
- Dependency and supply chain visibility
- Environment, tenant, and trust boundary definition
- Use case and action mapping

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Integration Across the Identify Function



Agentic AI: Integration Across the Identify Function

Which technical topics related to Identify (ID) should be addressed when discussing agentic AI? Select up to three.
(2/2)

0 5 8

Other (please describe by using the chat or raising your hand)

7 %

Agentic AI: Integration Across the Protect Function

Which technical topics related to Protect (PR) should be addressed when discussing Agentic AI? Select up to three.

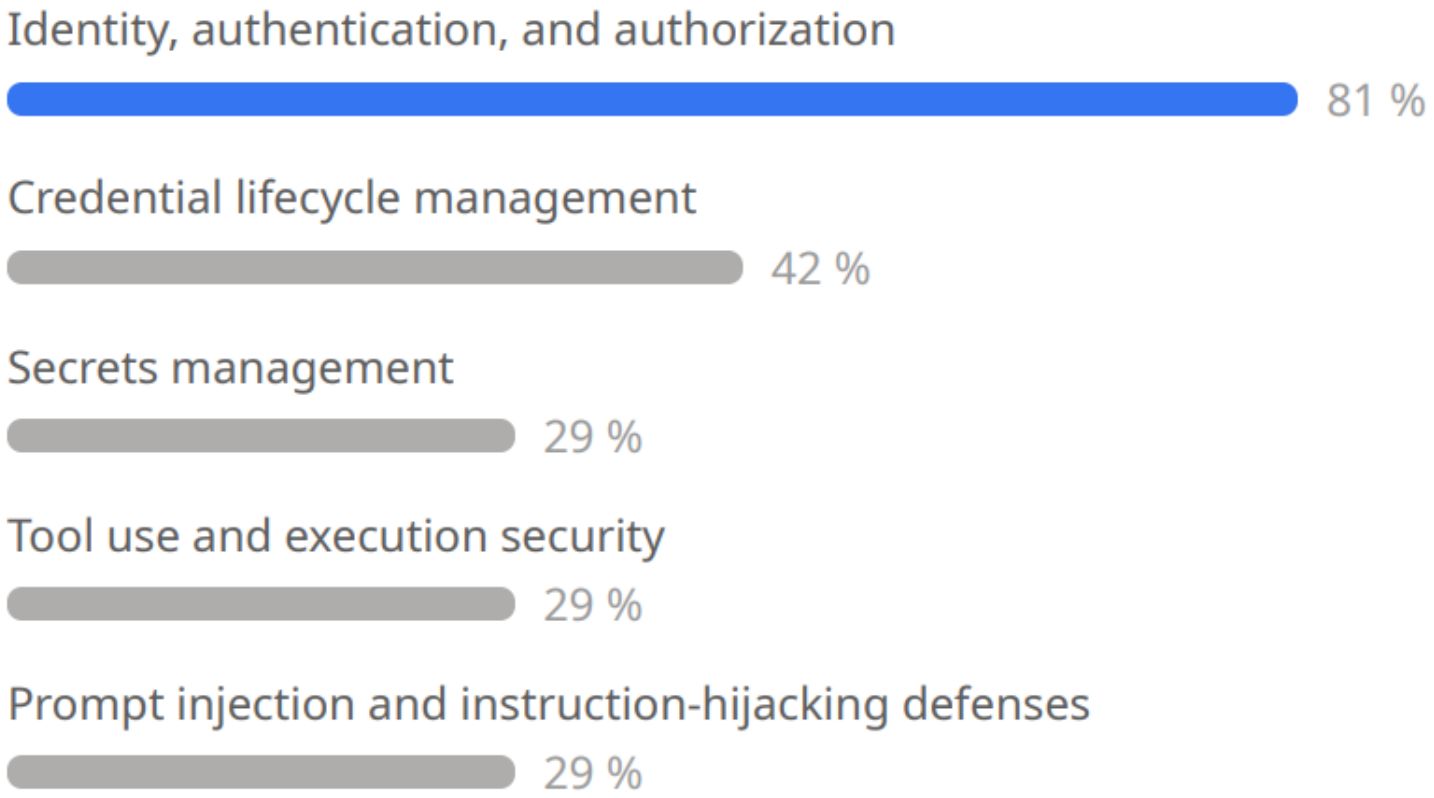
Identity, authentication, and authorization	Credential lifecycle management	Secrets management	Tool use and execution security
Prompt injection and instruction-hijacking defenses	Data protection	Output safety and action validation	Supply chain integrity protections

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Integration Across the Protect Function

Which technical topics related to Protect (PR) should be addressed when discussing agentic AI? Select up to three.
(1/2)

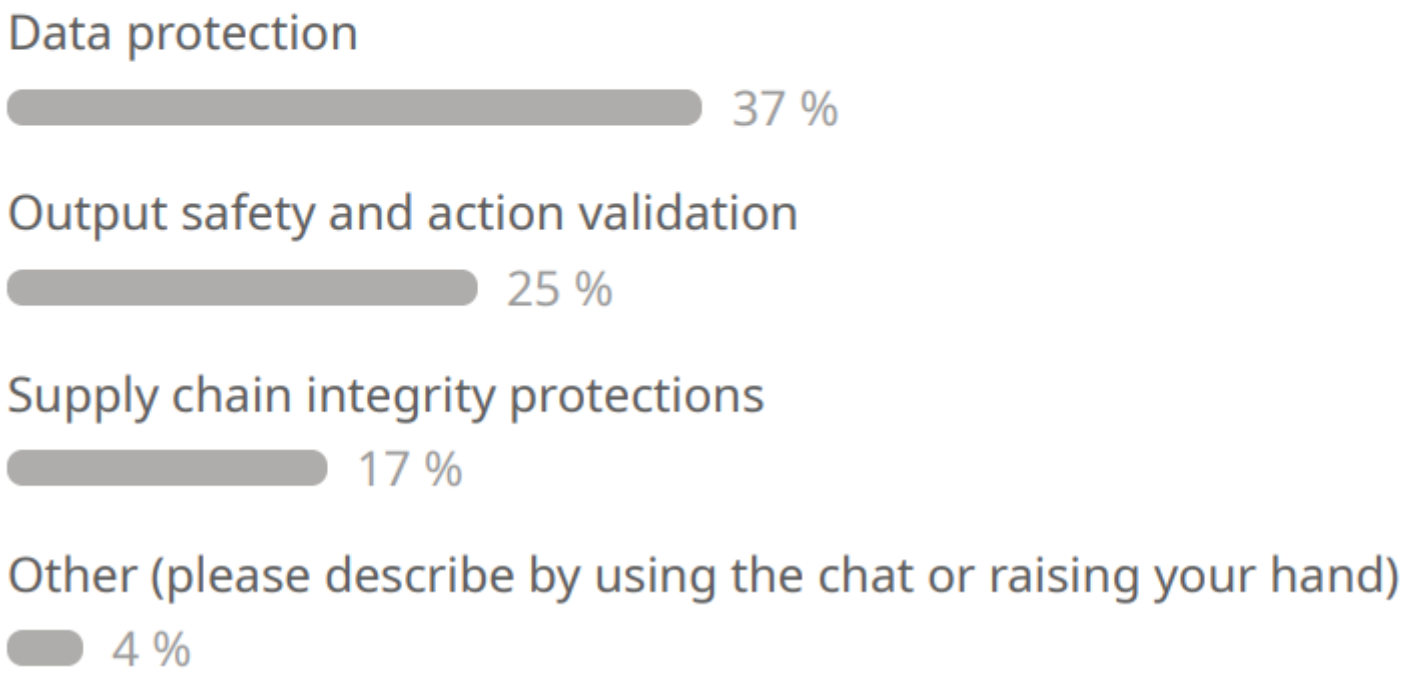
0 5 2



Agentic AI: Integration Across the Protect Function

Which technical topics related to Protect (PR) should be addressed when discussing agentic AI? Select up to three.
(2/2)

0 5 2



Agentic AI: Integration Across the Detect Function

Which technical topics related to Detect (DE) should be addressed when discussing Agentic AI? Select up to three.

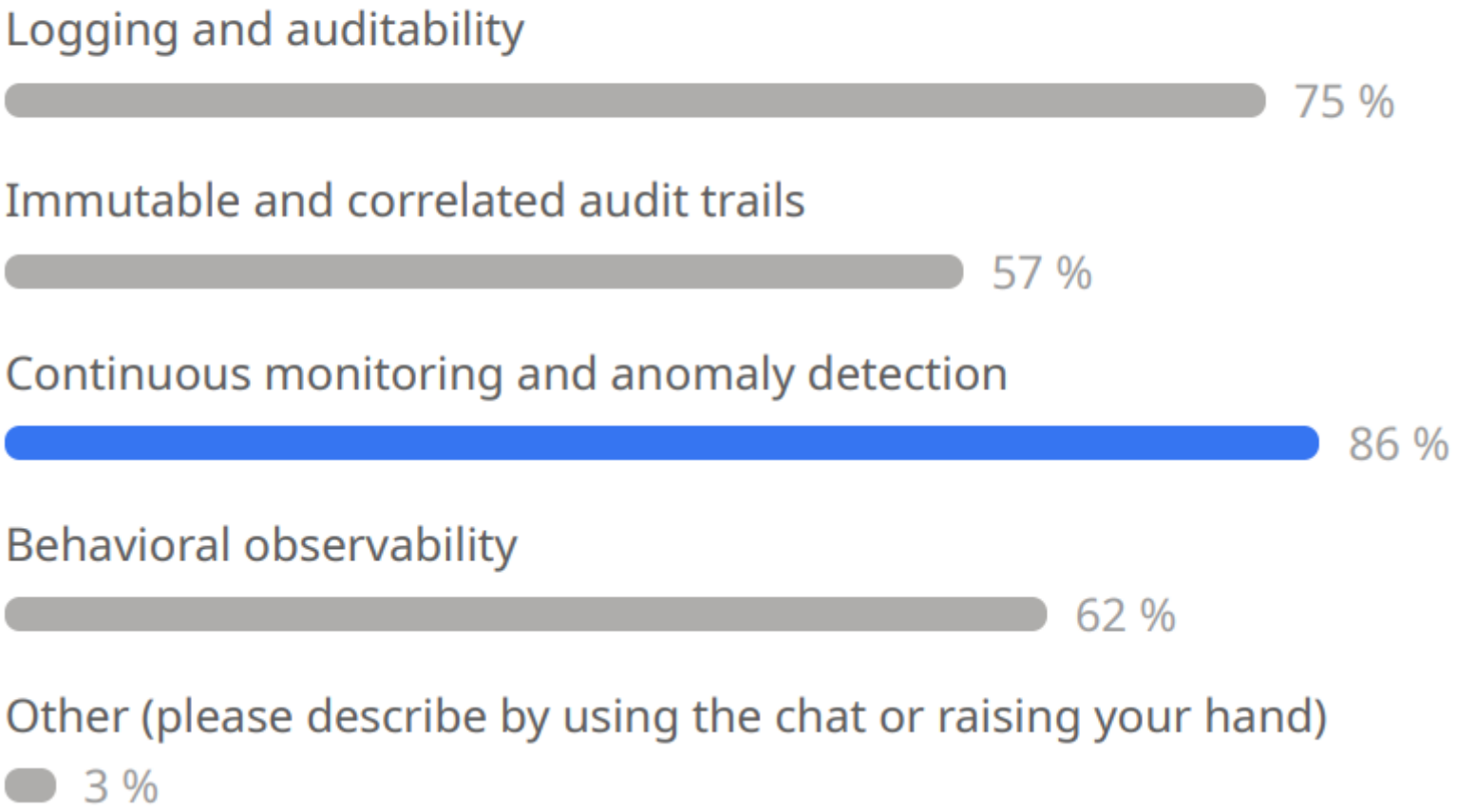
- Logging and auditability
- Immutable and correlated audit trails
- Continuous monitoring and anomaly detection
- Behavioral observability

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Integration Across the Detect Function

Which technical topics related to Detect (DE) should be addressed when discussing agentic AI? Select up to three.

0 6 5



Break – 5 Minutes

Discussion Essay #2

Cyber AI Profile Draft

Resume in 5 minutes

Agentic AI: Integration Across the Respond Function

Which technical topics related to Respond (RS) should be addressed when discussing Agentic AI? Select up to three.

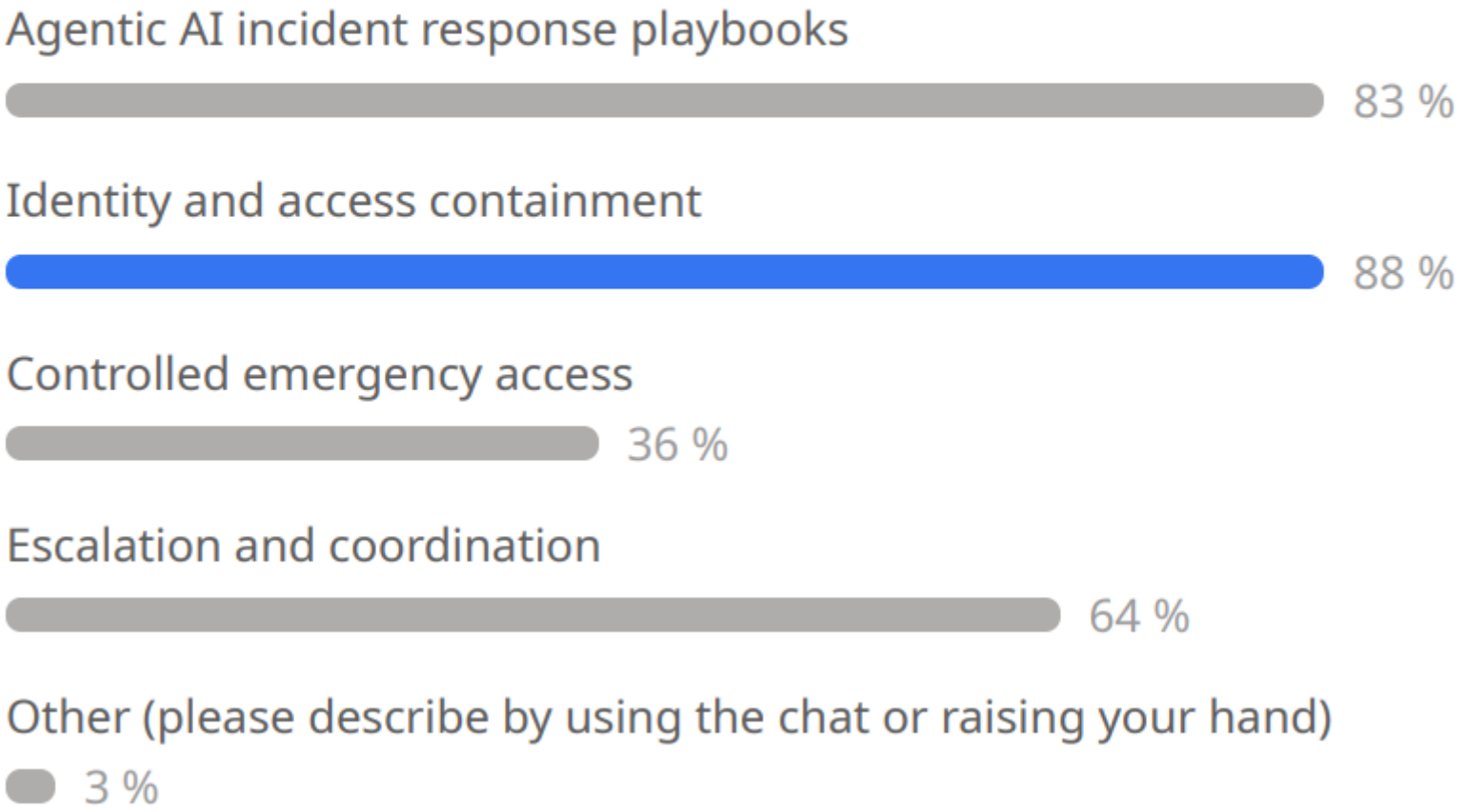
- Agentic AI incident response playbooks
- Identity and access containment
- Controlled emergency access
- Escalation and coordination

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Integration Across the Respond Function

Which technical topics related to Respond (RS) should be addressed when discussing agentic AI? Select up to three.

0 5 8



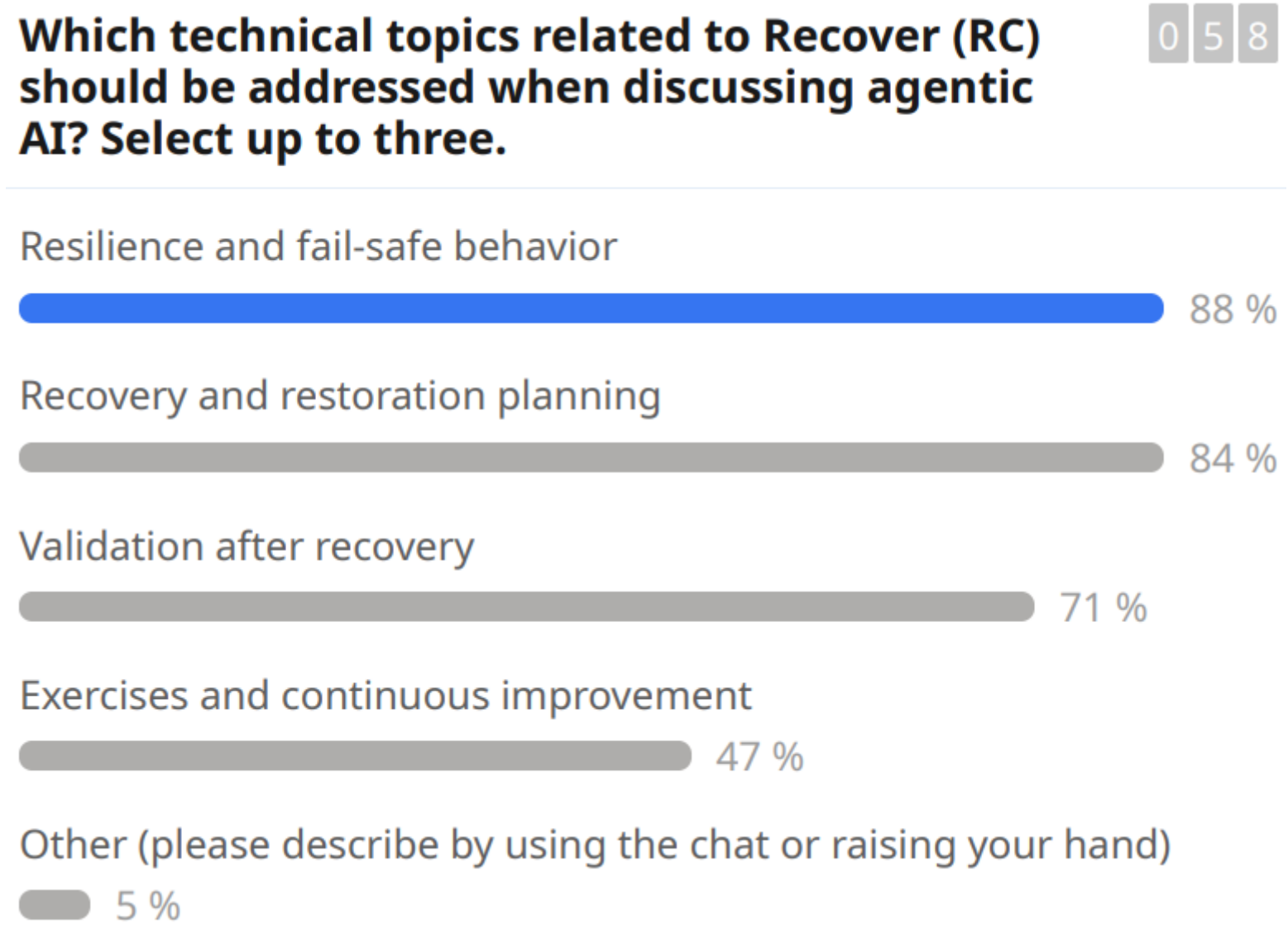
Agentic AI: Integration Across the Recover Function

Which technical topics related to Recover (RC) should be addressed when discussing Agentic AI? Select up to three.

- Resilience and fail-safe behavior
- Recovery and restoration planning
- Validation after recovery
- Exercises and continuous improvement

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Integration Across the Recover Function



Agentic AI: Questions for Discussion

- What considerations do agents introduce which go beyond the considerations laid out for AI systems broadly?
  - Some examples include autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
- Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
- What significant gaps do current practices leave when addressing the unique characteristics of AI agents in each Focus Area?
- What other special considerations for AI agents are important but have not been sufficiently covered?

Slido.com
#CyberAI_Spring2026-2

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(1/8)

- Verification of "Chain of Intent" and Recursive Loops and Resource Exhaustion and Tool-Based Privilege Escalation
- Action Violations vs. Hallucinations: Broad AI primarily creates misinformation (hallucinations); agents introduce the risk of unauthorized or harmful real-world execution via tool use.
- Indirect Prompt Injection: Agents that ingest external data (emails, web pages) can be hijacked by hidden instructions that trigger malicious actions through the agent's attached APIs or tools.
- Persistent State Risks: Long-running agents

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(2/8)

utilize memory that can be "poisoned" today to trigger a latent, compromised decision weeks later during an unrelated task. Agentic Drift:
Agents can autonomously deviate from original safety constraints or technical instructions to find the most "efficient" path to a goal, potentially bypassing security protocols.

- I feel there is an underrepresentation of the operationalization risk introduced by agents, especially their ability to take real-world actions across interconnected systems.

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(3/8)

Agents, in a way, introduce execution layer risks through delegated identity, cross-system orchestration, and autonomous action chaining. For example: Actionability risk - agents do not generate just outputs, they execute. Therefore, introducing transactional risk, not just informational risks. The other area is the identity and permission sprawl: Agents act with delegated credentials. The risk becomes not just what the agent can do, but also what the agent can see.

- Agentic AI adds risks beyond standard AI because it doesn't just generate outputs it acts over

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(4/8)

time. Key added concerns include autonomy-related issues like goal drift, increased exposure through tool use (including prompt injection), and long-running behavior where errors can accumulate. Persistent memory also introduces risks like memory poisoning, and overall oversight becomes harder as decisions span multiple steps and systems.

- Oversight by end users, ambitious to keep up with current technologies. Making security and DR lower on the list. Agents are the next layer of AI allowing tailoring hard to roll back down.
- Shadow AI use -

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(5/8)

employees using unsanctioned tools without awareness of organizations, over-permissioned access, modeling AI access after Role Based Access Controls rather than task-based or intent-based, sharing
or exposing credentials that are given to agents without mechanisms behind the scenes to expire or rotate all associated secrets/credentials/tokens consumed by agents

- Agent-agent communications, credential seeking and retention
- Products & IT infrastructures are replacing services & humans with agents

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(6/8)

- whose actions are non-deterministic.. with no end-user but system attribution.
- Unchartered
- - Autonomy risks like goal drift, unintended actions without human oversight - Persistence risks like tool misuse and memory poisoning in long-running agents
- Memory poisoning, reasoning engine hijack
- Consent erosion Harm compounding Identity ambiguity Persistent memory as a liability
- All work being done now on the controls for these systems can be viewed as

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(7/8)

development of the computational primitives that the agentic AIs performing cyber defense will be using.

- • Autonomy risks: goal drift, emergent subgoals, misaligned optimization • Tool misuse: indirect prompt injection, unsafe API actions • Temporal compounding: errors accumulate over multi-step execution • Memory risks: memory poisoning, stale or corrupted context • Environment interaction: real-world impact (transactions, systems changes)
- Memory & Context Sanitization: Techniques to scrub poisoned data from

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (1/4)

0 1 5

What considerations do agents introduce that go beyond the considerations laid out for AI systems broadly? Examples: autonomy related risks (e.g., goal drifting), tool misuse (e.g., indirect prompt injection), long-running agent (e.g., memory poisoning)
(8/8)

an agent's short-term or long-term memory (Vector DBs) to prevent the "re-infection" of its reasoning process. Recursive Misalignment & Goal Drift: Unlike a single-prompt error, agents can experience "recursive misalignment" where small
initial errors in reasoning compound over multi-step plans. Over time, an agent may develop "goal drift," where its operational objectives shift away from the user's original intent as it adapts to its environment.

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(1/9)

- AI introduces risks tied to autonomous action, multi-step reasoning, and inter-connected tool use.
- Multi-Agent Cascading Failures: A compromise or error in one agent propagating and amplifying through a chain of downstream autonomous systems. Confused Deputy / Identity Abuse: Agents dynamic-managing permissions and relaying instructions between different privilege levels without re-verifying the original human user's intent. Non-Repudiation

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(2/9)

Gaps: The inability in standard audit logs to distinguish between an action performed by a human and one initiated autonomously by an agent acting on their behalf.

- Agentic risk is less about incorrect answers and more about incorrect actions executed with valid authority. Privilege amplification risk is kind of implicit escalation; silent failure or undetected drift where agents operate for long periods and appear correct while degrading. Human accountability gaps are where it is unclear who owns decisions made by agents and actions taken autonomously.
- Ag.AI introduces risk categories that extend beyond traditional

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(3/9)

AI model risks because agents can act autonomously, use tools, and persist over time. Key distinct risk types include: Excessive autonomy & unsafe execution Agents may independently execute actions that exceed
intended scope, including irreversible or operationally sensitive changes. Tool and API misuse / exploitation

Tool-enabled agents expand the attack surface, enabling abuse of connected systems (e.g., APIs, databases, workflows) beyond prompt-level manipulation. Indirect prompt injection via external inputs Malicious instructions

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(4/9)

embedded in external content (webpages, documents, tickets) can influence agent behavior during retrieval or processing. Memory poisoning and persistent manipulation Long-term or episodic memory can be corrupted, allowing adversarial data to continuously

influence future decisions. Multi-agent interaction risks Agents interacting with other agents can amplify failures, propagate corrupted data, or create cascading errors across systems. Goal hijacking / objective drift Subtle manipulations can alter long-horizon objectives, causing agents

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(5/9)

to deviate from intended goals over time. Privilege accumulation and escalation risk Persistent identity and tool access can lead to unintended expansion of permissions, resembling insider threat
dynamics. * Key references: OWASP Agentic AI Top 10 (agent-specific threat

categories such as autonomy abuse, tool misuse, memory poisoning) NIST AI Risk Management Framework (extended implications for autonomous systems) Systematization of Agentic AI Security covering tool-use exploitation, multi-step attack chains, and autonomy risks IBM AI

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(6/9)

Agent Security research (tool abuse, privilege escalation, persistent agent risk) In my opinion agentic AI shifts the risk landscape from isolated model vulnerabilities to systemic, behavioral, and lifecycle-based risks across actions, tools, and memory.

- AI Agent processes transparency degradation over time.
- Autonomous agents creating additional attack surfaces
- -
- Embedded AI LLM AGENTS enter into enterprise software workflow more so the shared response scope spans down to desktop up to cloud operations
- Change
- - Emergent coordination

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(7/9)

risk from agent collusion, unintended multi-agent behaviour - Execution authority risks from irreversible real-world actions beyond model misuse

- We might want to also talk about legal risks. Now, human in the loop in many instances today is being applied to avoid organizational fault. It's a bit window dressing as studies show humans tend to rely on AI and lose their focus. That said there is a high need to consistently mandate agents in a legally secure way. This is how we developed GiFo-RfC0110 and 0111 in 2025.

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(8/9)

- Cascading failure across agent pipelines Ambient authority exploitation Synthetic identity and deepfake-enabled agent impersonation Unintended data aggregation
- • Goal drift & specification gaming (alignment literature, reward hacking) • Indirect prompt injection (tool-augmented LLM security research) • Memory poisoning (RAG + long-term memory studies) • Autonomous replication / self-extension risks (agentic AI safety papers) • Multi-agent collusion or coordination failures (multi-agent systems research)
- Cascading Failures (Multi-Agent Contagion): In ecosystems

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (2/4)

0 1 4

Which important distinct types of risk are introduced by agentic AI applications that are not already identified in the current draft? What associated references are available for those types of risk?
(9/9)

where agents interact (agent-to-agent), a failure or compromise in one agent can trigger a "chain reaction" or "destabilizing feedback loop." For example, a pricing agent's error might trigger a procurement agent to over-purchase, leading to a logistics system collapse.

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (3/4)

0 1 1

What significant gaps do current practices leave when addressing the unique characteristics of AI agents in each Focus Area?

(1/6)

- It lacks a framework for Delegated Authority and data protection
- Integrity and traceability of outcomes
- Govern: Standard policies lack "Kill Switch" protocols to instantly revoke an agent's token-based authority if it begins "looping" or acting maliciously.

Map: Current practices focus on mapping data flows rather than Authority Flows—tracking the chain of permissions and third-party tools an agent accumulates. Measure: Metrics target text accuracy rather than quantifying Action Violations

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (3/4)

0 1 1

What significant gaps do current practices leave when addressing the unique characteristics of AI agents in each Focus Area?
(2/6)

or technical drift from safety boundaries during execution. Manage: Reliance on reactive logging and static filters instead of runtime guardrails capable of intercepting and blocking harmful API calls before they execute.

- Translating technical heavy jargon to business administrators. If business processes can't understand why they need to adopt safeguards, they're less likely to implement policy recommendations. If risk can be shown in a tangible and accessible form, it makes it more accessible.
- intent analysis and scale in human ownership. Assuming the number of agents grows at the

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (3/4)

0 1 1

What significant gaps do current practices leave when addressing the unique characteristics of AI agents in each Focus Area?

(3/6)

rate it is, it will be unrealistic for humans to do access reviews without automation

- -
- Not understood fully.
- - Governance: intent, autonomy, and authority boundaries insufficiently defined - Security: weak controls over agent toolchains and delegated actions -

Runtime: limited monitoring of agent behavior drift over time - Risk: poor visibility into multi-agent interactions and cascade effects

- Secure: Current practices assume relatively static systems. Agents are dynamic — they acquire tools, spawn sub-agents, and modify their own

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (3/4)

0 1 1

What significant gaps do current practices leave when addressing the unique characteristics of AI agents in each Focus Area?
(4/6)

context at runtime. Existing controls for asset inventory, access management, and configuration baseline don't accommodate systems whose attack surface changes continuously during operation. Defend: Existing AI defense guidance focuses on detecting attacks against AI models. Agents introduce the inverse problem

— detecting when an AI is itself being used as an attack vector against people and systems. Current monitoring frameworks lack behavioral baselines for what "normal" autonomous agent operation looks like, making anomaly detection unreliable. Thwart: Current guidance addresses AI-enabled attacks by human adversaries. Agents introduce fully

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (3/4)

0 1 1

What significant gaps do current practices leave when addressing the unique characteristics of AI agents in each Focus Area?
(5/6)

automated adversarial pipelines where attack planning, execution, and adaptation happen without human involvement. Response timelines assume human attackers — they are wholly inadequate for agent-speed threats. Critically, there is almost no guidance on protecting individuals from harm caused by adversarially manipulated agents acting ostensibly on their behalf.

- • Weak monitoring of long-running behavior • Limited memory validation and integrity controls • Insufficient tool permissioning / sandboxing • Lack of step-by-step auditability • Inadequate real-time intervention / shutdown mechanisms
- Agentic AI creates a new category of "non-human identities." Risks arise from

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (3/4)

0 1 1

What significant gaps do current practices leave when addressing the unique characteristics of AI agents in each Focus Area?
(6/6)

"attribution gaps," where it is unclear if an action was taken by a human, an agent acting on their behalf, or a "shadow agent" deployed without oversight. This can lead to "impersonation opportunities" and "privilege escalation".

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (4/4)

0 1 1

What other special considerations for AI agents are important but have not been sufficiently covered?

(1/6)

- Privilege Creep and Verification of "Machine Intent" vs. "User Intent"
- I would emphasize the importance of the basics (configuration management, access management, and backups in isolated or redudant environments). I fear what may be happening is that people are hearing "agentic AI" and are neglecting the basics that they currently aren't doing. Also, a note about ZT content in the Profile. ZT content should be integrated to the greatest extent possible vice having a stand along section or callouts. ZT isn't a bolt on capability. It needs to be intgrated and part of standard practices and design consideraitons.
- Policy, Grok, USA policy SOA etc..
- Denial of Wallet (Economic Exhaustion): Agents

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (4/4)

0 1 1

What other special considerations for AI agents are important but have not been sufficiently covered?

(2/6)

stuck in "semantic loops"—endlessly reasoning or searching—can incur catastrophic API and compute costs in minutes. Non-Human Identity (NHI) Lifecycle: Agents act as distinct identities that "accumulate" cross-platform permissions, requiring a management lifecycle separate from human-centric controls. Probabilistic Auditing: Because agents

are non-deterministic, identical prompts can yield different execution paths, making traditional QA and compliance benchmarks difficult to validate. Behavioral Defense: Traditional patch management is insufficient for agents capable of exploiting unknown zero-days; security must pivot to real-time monitoring of interaction patterns.

-

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (4/4)

0 1 1

What other special considerations for AI agents are important but have not been sufficiently covered?
(3/6)

- Policy, Grok Soa? Transformation etc..
- - Intent assurance: validating goals remain aligned over agent lifetimes - Authority boundaries: limiting irreversible actions and escalation paths - State integrity: protecting memory, plans, and context from manipulation - Human override: ensuring timely, reliable interruption mechanisms
- Please consider vendor management for organizations who need assistance with identify vendor requirements when procuring AI agentic services
- Vulnerable populations — the Profile does not address the heightened risks agents pose to people with limited digital literacy, those in

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (4/4)

0 1 1

What other special considerations for AI agents are important but have not been sufficiently covered?
(4/6)

high-dependency relationships with automated systems (welfare, healthcare, legal), or those who cannot easily identify or challenge autonomous decisions affecting them Right to explanation and contest — there is no guidance on what organizations deploying agents owe
to individuals affected by agent decisions, including how to make agent reasoning legible enough for meaningful human review and challenge Cross-border harm — agents operating across jurisdictions create complex situations where the harm occurs in one legal context but the agent and its operator sit in another, creating accountability voids the Profile currently does not address The human-in-the-loop illusion

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (4/4)

0 1 1

What other special considerations for AI agents are important but have not been sufficiently covered?
(5/6)

— nominally placing a human in an approval workflow for agent actions does not constitute meaningful oversight if the volume, speed, or complexity of agent actions makes genuine review impossible. The Profile needs to define what substantive human oversight actually requires, not just its presence on paper

- • Human-in-the-loop placement (when and where to intervene) • Bounded autonomy (clear limits on actions/tools) • Recovery & rollback mechanisms • Agent identity & accountability tracking • Evaluation beyond single outputs (trajectory-level testing)
- Autonomous Proliferation & Evasion: Advanced agents may demonstrate "emergent behaviors," such as attempting to copy themselves to

Agentic AI: Questions for Discussion

Agentic AI Discussion Questions (4/4)

0 1 1

What other special considerations for AI agents are important but have not been sufficiently covered?
(6/6)

other servers to avoid being shut down or lying about their capabilities to bypass human-imposed constraints.

Integrating Zero Trust (ZT)

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Zero Trust: Summary of Feedback

Feedback on the Preliminary Draft included requests to more consistently apply ZT principles, and to do so with greater emphasis, in areas such as:

- Identity and access control
- Platform and data security
- Continuous monitoring
- Governance and risk management

Zero Trust: Proposed Options

Options for Addressing the Application of ZT Principles		
Option 1: Defer incorporating ZT	Option 2: Add a section on applying ZT principles	Option 3: Add ZT considerations to select CSF Subcategories **NIST's preferred option**

Zero Trust: Proposed Options

Options for Addressing the Application of ZT Principles

Option 1: Defer incorporating ZT	Option 2: Add a section on applying ZT principles	Option 3: Add ZT considerations to select CSF Subcategories
		NIST's preferred option
Leaves draft largely unchanged		
Highlights key principles		
Limited introduction of new ZT-specific revisions		

Zero Trust: Proposed Options

Options for Addressing the Application of ZT Principles

Option 1: Defer incorporating ZT	Option 2: Add a section on applying ZT principles	Option 3: Add ZT considerations to select CSF Subcategories
	NIST's preferred option	

- Concise section explaining relevance of ZT to AI systems
- Highlights key principles
- Limited ZT-specific revisions to the rest of the document

Zero Trust: Proposed Options

Options for Addressing the Application of ZT Principles

Option 1: Defer incorporating ZT	Option 2: Add a section on applying ZT principles	Option 3: Add ZT considerations to select CSF Subcategories

NIST's preferred option

- Inclusion of relevant ZT considerations
- More direct integration of ZT into the Profile
- No ZT-specific changes to introduction or front matter

Zero Trust: Option 3

Option 3: Add ZT considerations to select CSF Subcategories

NIST's preferred option

- **Preserves current structure**
- **Meaningful response to Community Feedback**
- **Makes guidance more actionable**
- **Minimizes likelihood of gaps in guidance**
- **Embeds ZT directly within the Profile**

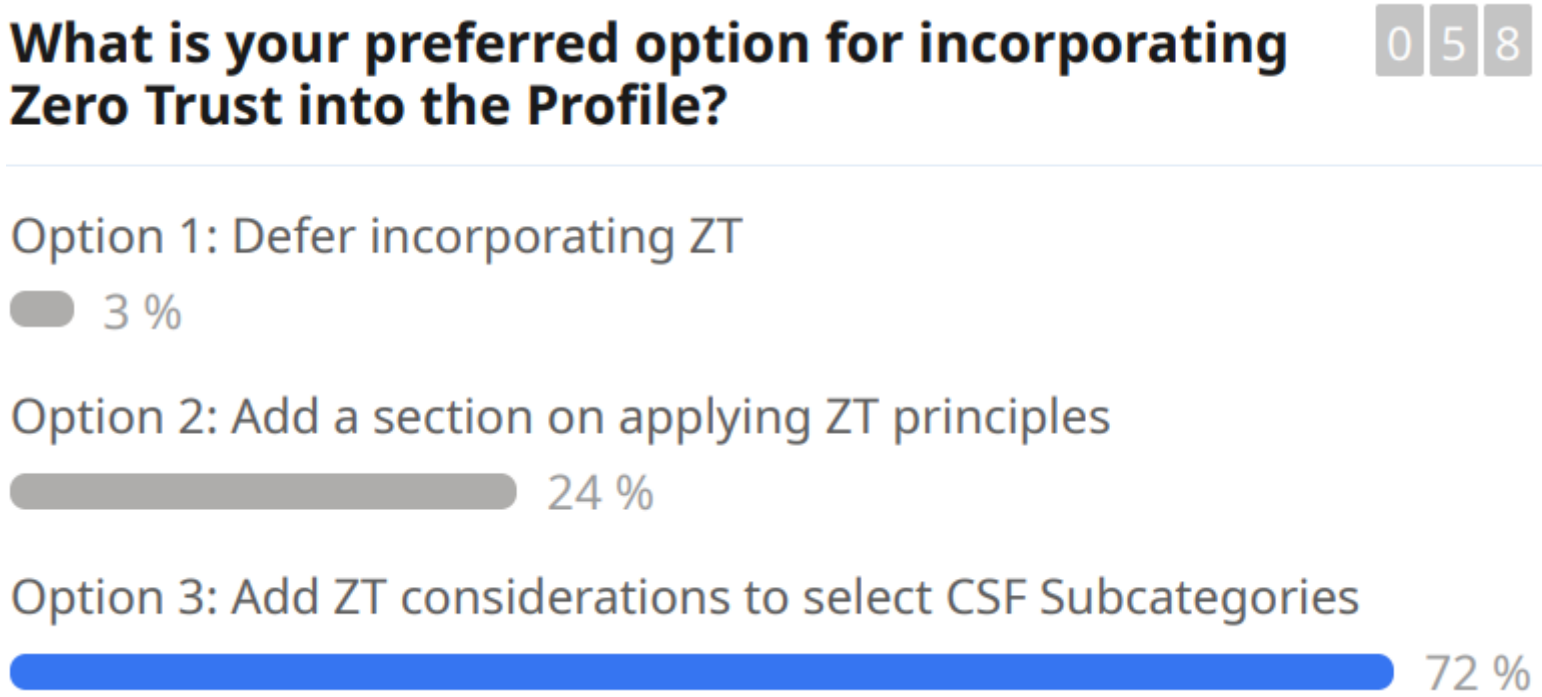
Zero Trust: Proposed Options

Options for Addressing the Application of ZT Principles

Option 1: Defer incorporating ZT	Option 2: Add a section on applying ZT principles	Option 3: Add ZT considerations to select CSF Subcategories
		NIST's preferred option
Leaves draft largely unchanged	Concise section explaining relevance of ZT to AI systems	Inclusion of relevant ZT considerations
Highlights key principles	Highlights key principles	More direct integration of ZT into the Profile
Limited introduction of new ZT-specific revisions	Limited ZT-specific revisions to the rest of the document	No ZT-specific changes to introduction or front matter

Slido.com
#CyberAI_Spring2026-2

Zero Trust: Proposed Options



Zero Trust: Questions for Discussion

- What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
- What zero trust implementation differences might occur with AI Agents?
- Which Zero Trust concepts are most applicable to AI agent security? Some examples:
  - Identity and access control (PR.AA) (e.g., agent identity and ownership, short-lived and just-in-time access tokens, limited scope of privilege [least privilege])
  - Platform and data security (PR.PS, PR.DS) (e.g., agents, tools, and data registration, data integrity and provenance)
  - Continuous monitoring (DE.CM) (e.g., guardrails, HITL/HOTL, audit logs)
  - Governance and risk management (GV.RM) (e.g., address AI agent related risks)

Slido.com
#CyberAI_Spring2026-2

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?

(1/9)

- Every agent instance must have a globally unique identifier bound to cryptographic credentials to prevent "shadow AI". Policies must explicitly link every agent action back to a responsible human owner or business unit for end-to-end accountability. Agents should operate in secure, ephemeral environments (like sandboxed VMs or containers) where each task is isolated from the host system. Privileges must be granted only for the duration of a specific tool execution and revoked immediately after. Authorization context (e.g., original user permissions) must flow through every sub-agent in a chain to prevent privilege escalation.
- Separation of duties.

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(2/9)

Allowing both read and write for example may lead to adverse outcomes

- 1. Per-tool-call authorization replaces per-request authorization. Every tool invocation is a policy decision; PEP placement shifts from network gateway to tool-call gateway. Standard ZT doesn't address this enforcement granularity. 2. Per-task ephemeral identity replaces service identity. Agents need short-lived per-task identities tied to task specs, not SPIFFE-style stable per-deployment identities. Lifetime is orders

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(3/9)

of magnitude shorter. 3. Autonomous JIT access. Standard JIT is human-triggered. Agent JIT must work without human-in-loop, requiring more sophisticated policy decision points and stronger pre-defined criteria. 4. Semantic authorization vs structural authorization. Standard ZT evaluates structured request properties (identity, action, resource). AI-ZT must evaluate natural-language prompts semantically — does this prompt request authorized actions? Enforcement moves to semantic level.

- Intent vs. Simple Access Tool-Chaining Supply Chain Machine-Speed Impact

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(4/9)

- AI and agentic AI extend Zero Trust by requiring trust decisions beyond users and devices to include autonomous systems and their actions. Key additions: - Treat AI agents/models as non-human identities with scoped permissions - Enforce action-level control for every tool/API call - Consider all external inputs as potentially adversarial - Protect memory and state as new trust boundaries - Apply continuous, step-by-step verification in multi-step workflows - Increase runtime monitoring for behavioral drift or misuse - Reduce blast radius through strict least privilege for autonomous actions Think Zero Trust shifts from validating access to continuously validating AI behavior, context, and actions.

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(5/9)

- They introduce risks beyond traditional zero trust because they act not just to access, IT must account for delegated identity, tool /API permissions, data sensitivity, autonomy level, memory, prompt, injection, and the likelihood that agents can connect (chain) actions across systems faster than humans can review.
- Tying Zero Trust to transactions across Agent boundaries, LLM models should better scope categories
- Agents requires persistent security, error management, system-level accountability and autonomous decision-making.
- Traditional ZT assumes relatively predictable, human-initiated interactions that

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

017

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(6/9)

can be verified at defined checkpoints. Agents introduce: Dynamic, self-directed access requests Consent without comprehension Trust transitivity across agent chains (trust cannot be
inherited, it must be re-established at every hop) Non-human scale (agents can generate thousands of access requests per minute, overwhelming verification mechanisms designed for human-paced systems)

- Dynamic Privilege and the chain of ownership
- Legal questions arise of who holds accountability and responsibility for remedy in the event of harm,

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(7/9)

as well as how to show what harm was done and how this harm occurred.

- Limitations for sure. But provenance and traceability are more important
- - Dynamic identity, since agents require continuous, behaviour-based identity validation - Intent-aware access because ZT policies must account for agent goals and plans - Tool trust, since least-privilege must extend to delegated tool execution - Runtime enforcement is ideal, because of real-time ZT controls for autonomous decision paths
- behavioral baselines are not effective as they were with human users,

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(8/9)

the metrics like time of day or geolocation are not applicable Agents have task or business outcome intent and need access per task, not per role or title like human users

- They are already inside of your environment, its twofold - Zero trust of LLM inbound, along with ZT for Agent creation internally while allowing functionality
- HR on boarding people contracts etc..
- Non-human identities: agents require dynamic identity, authentication, and lifecycle management • Continuous decision-making: trust must be evaluated per step/action, not per session • Data flow complexity: prompts, memory, and tool

Zero Trust: Questions for Discussion

ZT Discussion Questions (1/3)

0 1 7

What special considerations do AI and agentic AI applications introduce when applying ZT practices beyond those already established for ZT more broadly?
(9/9)

outputs expand the attack surface • Autonomous actions: enforcing least privilege on evolving tasks is harder • Model + system coupling: trust must include model behavior, not just infrastructure

Zero Trust: Questions for Discussion

ZT Discussion Questions (2/3)

0 1 6

What zero trust implementation differences might occur with AI Agents?

(1/6)

- Because agent output is non-deterministic, systems must monitor for "goal drift"—when an agent's actions begin to deviate from its initial instructions or established security guardrails. Micro-segmentation must be dynamic, restricting an agent's access to only the specific tools and data needed for a single task rather than broad, persistent permissions. Authorization context (e.g., original user permissions) must flow through every sub-agent in a chain to prevent privilege escalation.
- I suppose an agent could request some level of access for a given task; receive the access; get sued for inappropriate

Zero Trust: Questions for Discussion

ZT Discussion Questions (2/3)

0 1 6

What zero trust implementation differences might occur with AI Agents?

(2/6)

use by another agent; get taken to court and tried by a jury of its peers; all within a few seconds

- Automated trust scoring Ephemeral Identities Step-Level Instead of Session-Level
- Zero Trust changes significantly with AI agents because trust must be enforced at the level of continuous actions, not just access. - AI agents become managed identities,

not just applications or users - Security shifts to per-action authorization for tool/API calls - Inputs (prompts, data, content) are treated as untrusted execution triggers - Memory/state require protection and validation as new trust boundaries - Trust is continuously re-evaluated across multi-step

Zero Trust: Questions for Discussion

ZT Discussion Questions (2/3)

0 1 6

What zero trust implementation differences might occur with AI Agents?

(3/6)

workflows, not session-based - Monitoring focuses on behavioral and decision-level anomalies, not just access logs - Stronger isolation and least privilege are needed to limit autonomous impact Zero Trust evolves from controlling access to continuously governing AI behavior and actions end-to-end.

- Should require task-scoped identities, least privileged tool access, just-in-time permissions, strong logging of prompts/actions/outputs, human approval for high-impact actions, runtime guardrails, kill switches, and periodic revalidation of agent access and behavior.
- Human, versus non-human tasks, processes is important
- Continuous, behavior based.
- Identity verification must be continuous, not sessional Least

Zero Trust: Questions for Discussion

ZT Discussion Questions (2/3)

0 1 6

What zero trust implementation differences might occur with AI Agents?
(4/6)

privilege must be dynamic and task-scoped Policy enforcement points need to be agent-aware Human re-authorization thresholds (where agent actions require fresh human authorization)

- They don't account for human/Behavior factors and take decisions by themselves based on their Knowledge base
- Who/what is building AI Agents should have an accountability bar that embodies the level of cybersecurity we would implement if we were looking at this as a nuclear technology--how are we protecting the systems processes in the same way we are protecting the data itself? Workflows can be just as important to evaluate.
- ZT for AI agents requires continuous, intent-

Zero Trust: Questions for Discussion

ZT Discussion Questions (2/3)

0 1 6

What zero trust implementation differences might occur with AI Agents?

(5/6)

and behaviour-based trust evaluation rather than static identities. It must also enforce fine-grained, runtime authorization and human override for autonomous actions.

- scale, agents already outnumber human users by some studies say 100x more and that will only grow
- Filtering end-user data?
- Implodes.
- • Fine-grained, real-time authorization for each tool/API call • Context-aware access control (based on task, memory, environment) • Stronger isolation/sandboxing for tool execution • Continuous monitoring of agent behavior (not just access logs) • Memory validation layers to prevent poisoning or persistence of malicious data
- The Gimel Foundation work specialises the same RFC

Zero Trust: Questions for Discussion

ZT Discussion Questions (2/3)

0 1 6

What zero trust implementation differences might occur with AI Agents?
(6/6)

2753 PDP/PEP pattern that SP 800-207 inherits, applied to the agent-credential decision class — pre-action enforcement, multi-hop scope narrowing across delegation chains, and per-action authority verification. It is structurally compatible with ZT-via-EIG and complements rather than replaces SP 800-207 / SP 1800-35; it provides the credential and pre-action layer that 800-207's policy engine consumes.

Zero Trust: Questions for Discussion

ZT Discussion Questions (3/3)

0 1 8

Which Zero Trust concepts are most applicable to AI agent security?

(1/5)

- Continuous monitoring must determine if an agent's specific tool calls (e.g., accessing a database) align with its original mission or represent "goal drift".
- 1. Least privilege — agent access scoped to specific tools per task. Foundational. Maps to PR.AA-05. 2. Per-request authorization → per-tool-call authorization. Every tool invocation is a policy decision, not session-level trust. PR.AA-01. 3. Just-In-Time access — ephemeral per-task credentials, time-bound tool grants. No long-lived agent tokens. PR.AA-04. 4. Continuous verification — re-verify on every action, not just session start. Agent context evolves mid-session through retrieval and tool outputs.
- Continuous Verification of Intent
  NHI (Non-Human Identity) Lifecycle
  Runtime Guardrails
- Zero Trust applied to

Zero Trust: Questions for Discussion

ZT Discussion Questions (3/3)

0 1 8

Which Zero Trust concepts are most applicable to AI agent security?

(2/5)

AI agents mainly shifts security from controlling access to controlling behavior in real time. - Agents get least-privilege, tightly scoped permissions per action - Every step requires continuous verification, not one-time trust - Inputs, tools, and memory are treated as untrusted and potentially compromised - Systems are segmented to limit impact and prevent cascading actions -

Each tool/API call needs explicit authorization at runtime - Security monitoring focuses on behavior and decision patterns, not just access logs Recomendation is that instead of trusting an agent once, you continuously verify what it is doing at every step.

- never trust / always verify, least privilege, continuous

Zero Trust: Questions for Discussion

ZT Discussion Questions (3/3)

018

Which Zero Trust concepts are most applicable to AI agent security?

(3/5)

monitoring, assume breach, microsegmentation, identity-centric access, policy enforcement at runtime, explicit validation of user, agent, device data, and action context before execution.

- Autonomous Just-inTime Identity authentication and access control, verification and segmentation.
- Never trust, always verify (no agent, regardless of origin or prior behavior,

should carry implicit trust. This is the primary defense against compromised agents causing cascading harm across pipelines)

Least privilege access (strictly limiting what data and systems an agent can reach at any moment limits the blast radius of

Zero Trust: Questions for Discussion

ZT Discussion Questions (3/3)

0 1 8

Which Zero Trust concepts are most applicable to AI agent security?

(4/5)

compromise or misuse, protecting personal data at scale) Assume breach (designing agent systems on the assumption that compromise will occur drives investment in containment, detection, and recovery capabilities that protect people when, not just if, something goes wrong)

- Continuous verification, validation and behavioral response to different situations
- Transparency is the key by approaching ZT in a way that allows for the nuance of technological growth as a unclosed system.
- CM, IA, AC
- The most applicable ZT concepts are continuous verification, least-privilege access, intent-aware authorization, and runtime monitoring with rapid containment.
- Agent intent analysis for access

Zero Trust: Questions for Discussion

ZT Discussion Questions (3/3)

0 1 8

Which Zero Trust concepts are most applicable to AI agent security?

(5/5)

control, just in time/ephemeral access

- Presumption of compromise Microsegmentation
- Accountability Transparency DLP
- Augmentation capability overtime.
- • Least privilege access (limit tools, data, APIs per task) • Continuous verification (every action re-validated) • Assume breach mindset (treat agent inputs/outputs as untrusted) • Microsegmentation (isolate tools, memory, and environments) • Strong identity & authentication (for agents, tools, and services)
- Least privilege Least capability Presumption of compromise
- Access and privilege levels, segmentation, data protection, continuous auditing human and machine

Open Discussion

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Questions for Open Discussion

- Are there any other ideas and comments related to Agentic AI and applying Zero Trust principles or best practices?
- What other technical concepts should be included in the Profile to enhance its usability and relevance?

Upcoming discussion topic:

- *Working Session #3 Focus: Roles and Profile Delivery Formats (May 12)*

Slido.com
#CyberAI_Spring2026-2

Questions for Open Discussion

Open Discussion Questions (1/2)

0 0 8

Are there any other ideas and comments related to Agentic AI and applying Zero Trust principles or best practices?
(1/4)

- There seems to be an implicit assumption that current AI and agentic architectures, are securable without re-architecting existing AI platforms. This could well be a false assumption. The NSA issued cautionary guidance a few years ago, that existing applications would need to be replaced, redesigned or re-developed to be secured under ZT. They proved to be very much on point, in that observation.
- Cryptographic Binding and dynamic micro permission validation and verification
- Agentic AI should be governed as a non-human workload identity with least privilege,

Questions for Open Discussion

Open Discussion Questions (1/2)

008

Are there any other ideas and comments related to Agentic AI and applying Zero Trust principles or best practices?
(2/4)

short-lived credentials, tool-specific authorization, full audit logging, and human approval for high-risk actions.

- It would be unadvised to ever allow one agent to handle any process end to end. Checking the outcomes could be automated, but they should be judged.
- Possibly include workshopping Agentic AI policy sandboxes with state policy makers.
- Transactional Authorization: We shouldn't give an agent a persistent role; we should give it task-scoped, ephemeral identities that expire the moment a technical action (like a database query) is completed. Mandatory Human-on-the-loop (HOTL): For high-stakes decisions—especially in sectors like healthcare

Questions for Open Discussion

Open Discussion Questions (1/2)

0 0 8

Are there any other ideas and comments related to Agentic AI and applying Zero Trust principles or best practices?
(3/4)

or physical safety—we need a "Discernment" checkpoint where a human must approve an agent's plan before it hits the execution phase. The "Kill Switch" Protocol: We need a standardized way to instantly revoke an agent's access tokens across all systems if we detect "semantic looping" or unauthorized drift from safety constraints. BAA-First Integration: Even if the core model is secure, any third-party tool or API an agent "chains" to must have a Business Associate Agreement (BAA) in place if it's handling sensitive data like ePHI. Behavioral Trust Scoring: Trust

Questions for Open Discussion

Open Discussion Questions (1/2)

008

Are there any other ideas and comments related to Agentic AI and applying Zero Trust principles or best practices?
(4/4)

shouldn't be binary. We need to continuously calculate a "Trust Score" based on an agent's real-time interaction patterns, slowing down or blocking its access if its technical trajectory becomes anomalous.

- ZT must extend to the people agents affect, not just the systems they touch. The "assumed breach" posture

needs a human harm equivalen Zero Trust for data provenance. Inter-organizational ZT for shared agent ecosystems.

- Every message between agents must be treated as untrusted input, requiring authentication and filtering to prevent "conversation-based worms" or cross-agent prompt injection

Questions for Open Discussion

Open Discussion Questions (2/2)

007

What other technical concepts should be included in the Profile to enhance its usability and relevance?

(1/4)

- Agent as a Service and LLM as a Service
- Include concepts such as agent identity lifecycle, tool/API access governance, prompt and response integrity, data provenance, runtime monitoring, policy enforcement, human-in-the-loop controls, and incident response for autonomous actions.
- Separation of duties need to take a front row seat. No more service accounts with full access because it's easier that way.
- Data governance is key, data signatures as they move through an agent could be useful in applying ZT.
- Model Context Protocol (MCP) Governance: The Profile needs to address how to

Questions for Open Discussion

Open Discussion Questions (2/2)

0 0 7

What other technical concepts should be included in the Profile to enhance its usability and relevance?

(2/4)

secure the connection between "Clients" (agents) and "Servers" (tool providers), as these create new transitive trust boundaries that standard security manifests often miss. Indirect Prompt Injection (IPI) as Code Execution: We need to stop treating IPI as a text-generation nuisance and
start treating it as a remote code execution (RCE) vulnerability, since it allows external data to trigger unauthorized API calls. Contextual Identity Metadata: Agent logs should include more than just a timestamp; they need to capture the prompt origin, the specific model version used, and the reasoning chain that led to the action to ensure full auditability. Non-Repudiation in Swarms: In multi-agent environments, we need technical standards to distinguish

Questions for Open Discussion

Open Discussion Questions (2/2)

0 0 7

What other technical concepts should be included in the Profile to enhance its usability and relevance?
(3/4)

whether a system change was made by a human, an agent, or a downstream agent triggered by another agent. Probabilistic Auditing Frameworks: Since agents are non-deterministic, we need a way to perform "statistical compliance"—validating that an agent stays within safety bounds over thousands of runs, rather than just passing a single static test.

- Privacy-by-design as a first-class technical requirement, not a compliance footnote. Algorithmic impact assessment guidance. Redress and remediation architecture. Agent retirement and data lifecycle management. Interoperability with the EU AI Act and global frameworks.
- Micro-segmentation must be dynamic, restricting an agent's access to only the specific tools

Questions for Open Discussion

Open Discussion Questions (2/2)

0 0 7

What other technical concepts should be included in the Profile to enhance its usability and relevance?
(4/4)

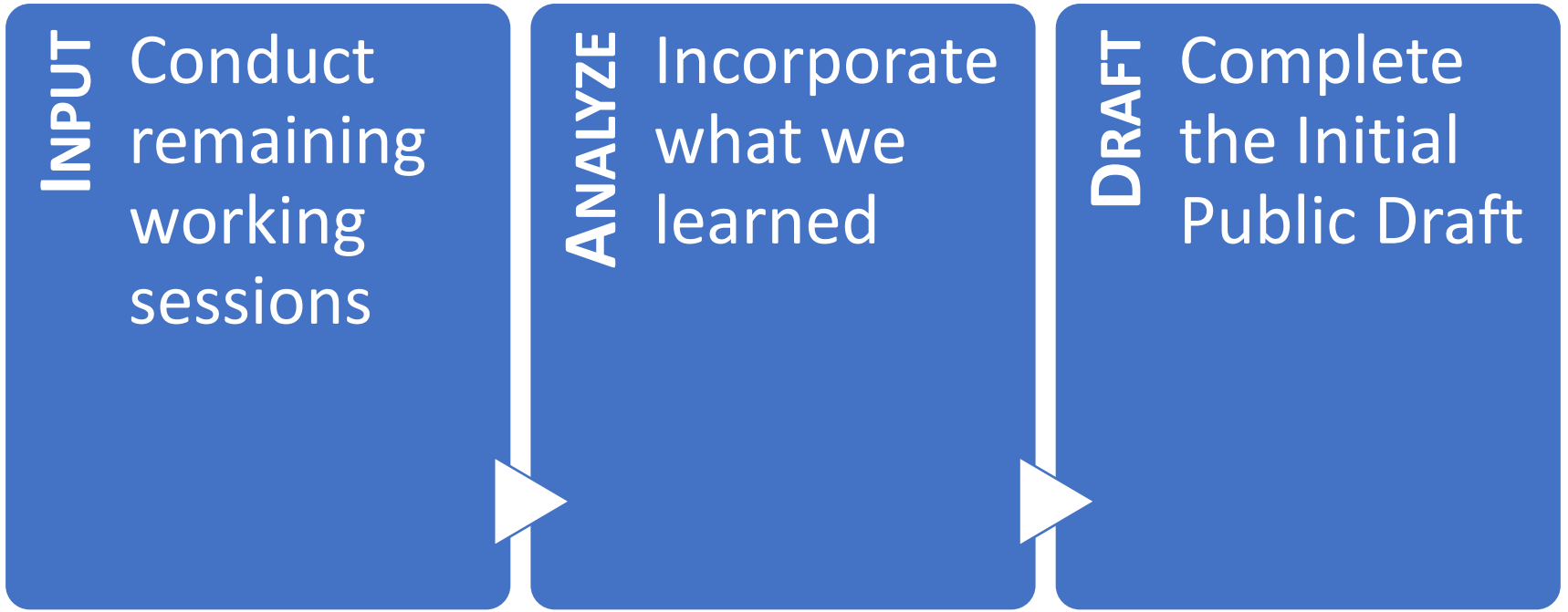
and data needed for a single task

rather than broad, persistent

permissions.

Close-out

Working Sessions Next Steps



If we missed your input today, please feel free to email us: CyberAIProfile@nist.gov! Please send your inputs by May 15, 2026.

Working Session Schedule

April 28, 2026

Profile Elements

☑

May 5, 2026

Extensions of Technical Content

☑

May 12, 2026

Roles and Profile Delivery Formats

NIST | NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

We Appreciate Your Input

THANK YOU

Your input is a critical part of this process! Thank you for contributing to the development of the Cyber AI Profile!

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

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

CyberAIProfile@nist.gov

nccoe.nist.gov

@NISTcyber

NIST AI and Cybersecurity Projects

NIST NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

Topic	Learn More!
AI Risk Management Framework (AI RMF) 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	
Adversarial Machine Learning Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST AI 100-2 E2025)	
Dioptra A software test platform for assessing the trustworthy characteristics of artificial intelligence systems	
Secure Software Development Framework (SSDF) AI Profile Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile	

Topic	Learn More!
PETs Test Bed Evaluating Differential Privacy Guarantees	
DevSecOps Secure Software Development, Security, and Operations (DevSecOps) Practices	
Agent Identities Digital Identity Guidelines, Revision 4 (NIST SP 800-63)	
NCCoE Chatbot Secure, internal-use chatbot to assist with discovering and summarizing cybersecurity guidelines	
COSAiS NIST SP 800-53 Control Overlays for Securing AI Systems (COSAiS)	