Cybersecurity and AI Workshop Concept Paper

National Institute of Standards and Technology

February 14, 2025

Note to Reviewers

Recent advancements in Artificial Intelligence (AI) technology bring great opportunities and challenges to organizations, including how AI can affect their cybersecurity capabilities and risks. The potential positive and negative impacts of AI need to be understood and managed. Long active in both cybersecurity and AI, NIST has worked closely with stakeholders in conducting research and providing guidance. The agency is exploring how to best use its cybersecurity resources to assist organizations as AI presents new opportunities and risks.

NIST's extensive engagement with cybersecurity stakeholders has led to several initial observations:

- There is no consistent taxonomy or agreement on how AI advances inform organizations' strategies for cybersecurity risk management.
- Cybersecurity professionals must strategically address emerging cybersecurity risks stemming from advancements in AI, even as they continue to manage ongoing operations.
- These professionals would benefit from informed, neutral guidance and other resources to inform their strategies and help them to organize and prioritize their actions.
- AI introduces new challenges with potentially major impacts regarding cybersecurity – but AI advances do not necessarily require fundamental changes to the way organizations address cybersecurity. Existing cybersecurity standards, frameworks, guides, and practices can still be effective when used individually and together if they are applied or modified to specifically address AI-related challenges as well as AI's helpful capabilities.

NIST welcomes stakeholder feedback to inform our path forward as we seek to better understand cybersecurity and Ai considerations and provide practical approaches to help address them. For example, many cybersecurity professionals have told NIST that a profile of the NIST Cybersecurity Framework (CSF) 2.0 may be one practical way to help organizations to understand, examine, and address cybersecurity risks – and opportunities – introduced by the adoption of AI.

Another suggestion NIST has heard is to provide a community profile of the NIST AI Risk Management Framework (AI RMF) or multiple crosswalks of that document with various cybersecurity resources. (The AI RMF identifies "Secure and Resilient" as one of the primary

characteristics of AI trustworthiness.) Other approaches may exist, and the best path forward might involve pursuing multiple options.

As part of its broader goal of assessing how NIST can best assist organizations and professionals at the intersection of cybersecurity and AI, NIST will host its first Cybersecurity and AI Workshop on April 3, 2025, to inform next steps. As input to the upcoming workshop, NIST's National Cybersecurity Center of Excellence (NCCoE) has prepared a draft concept paper to inform development of a possible CSF Community Profile based on the intersection of cybersecurity and AI. NIST seeks feedback on whether this profile or other resources would help organizations better understand and manage cybersecurity risks related to AI development and use.

In particular, NIST is interested in stakeholder perspectives on the following questions:

- **Scope:**
  - Is it appropriate to develop or modify existing cybersecurity-focused guidelines and resources to specifically address how AI advances change cybersecurity risks and opportunities?
  - Is it appropriate to separately develop or modify existing AI-focused risk management guidelines and resources to specifically address cybersecurity considerations?
  - Are we focusing on the right areas (securing AI system components, thwarting AI-enabled attacks, and leveraging AI in organizations' cybersecurity approaches)?
  - Are there any key areas missing that are at the intersection of cybersecurity and AI?
  - Should AI design and implementation failures be included?
- **Cybersecurity Risks:**
  - What existing NIST guidance or best practices should be included to address the needs of various stakeholders?
  - What gaps in NIST guidance exist that should be filled?
- **Multi-dimensional Views:**
  - Should NIST expand this effort to include demonstrating the relationship between cybersecurity and privacy of AI?
  - In what ways might NIST better represent the relationship across its cybersecurity, privacy, and AI resources?
- **Related Efforts:**
  - What groups and activities should we connect with to inform our efforts?
  - Are there emerging standards activities we should consider? If so, which ones?

- **Other Considerations:**
    - What else should we consider to inform the development of meaningful guidance in these areas?

Feel free to share your thoughts with us via CyberAIProfile@nist.gov by March 14, 2025. Your feedback will help drive the agenda and discussion during the upcoming Cyber and AI Workshop. More details forthcoming!

1. Introduction

Recent advances in Artificial Intelligence (AI) technologies bring great opportunities to organizations, but also new risks and impacts that need to be managed in the domain of cybersecurity. The National Institute of Standards and Technology (NIST) is evaluating how to use existing frameworks, such as the NIST Cybersecurity Framework (CSF), to assist organizations as they face these new or expanded risks. Discussions with many in the cybersecurity community strongly suggest that there would be value in developing guidance based on the CSF to address the cybersecurity risks related to AI development and use.

Organizations vary on whether and how they are using AI within their operations. Some organizations may not yet be using AI. Some may be using cybersecurity solutions enabled with basic machine learning (ML) technology for pattern detection and planning but have not yet adopted newer and transformational AI capabilities, such as Generative AI. Regardless of where organizations are on their AI journey, they need risk management approaches that support the realities of advancements in AI use to position them for defending against AI-enabled cyber offense by adversaries and taking advantage of AI-enabled cyber defense capabilities.

2. Cybersecurity of AI and AI in Cybersecurity (Cyber AI) Profile Conceptual Approach

The outcomes described in the NIST Cybersecurity Framework (CSF) 2.0 are one practical way to help organizations understand, examine, and address the cybersecurity risks introduced by the adoption of AI. The NIST CSF is used to understand, assess, prioritize, and communicate cybersecurity efforts. A CSF Community Profile describes CSF outcomes to address shared interests and goals among multiple organizations, and it provides guidance organized around a specific sector, technology, threat type, or other use case. Community Profiles are helpful for describing cybersecurity risk management priorities of a community, encouraging common target outcomes, aligning requirements from multiple sources, and leveraging expertise across the community to the benefit of all organizations in that community. Examples of Community Profiles that have been previously developed for various sector, technology, and cybersecurity use cases can be found on the NCCoE's Framework Resource Center.[1] In this case, a Cyber AI Profile could provide guidance for organizations that are deploying AI technologies and/or defending against AI-enabled attacks.

Examples of AI-related sources of cybersecurity risk that can impact an organization's operational risk include: Cybersecurity of AI Systems, AI-enabled Cyber Attacks, and AI-enabled Cyber Defense. The Cyber AI Profile could organize the discussion of these cybersecurity implications around the Functions, Categories, and Subcategories of the CSF Core. There is already extensive work in the public, private, and academic sectors in many of these areas; the Cyber AI Profile could highlight existing material as illustrative examples and Informative References useful to organizations as they endeavor to update their cybersecurity risk management approaches.

[1] For more information about creating and using Community Profiles, see the NCCoE Framework Resource Center's Guide to Creating Community Profiles.

Consider how the CSF Categories and Subcategories in Examples #1-3 below can be applied to address cybersecurity risks from AI. These are some of the potential implications for cybersecurity risk management activities which organizations may need to update to keep pace with AI development and use.

Example #1:

Awareness and Training (PR.AT) With advancements in the use of AI by cyber adversaries, organizations may need to consider revising their employee training (PR.AT-01) to raise awareness of AI-enabled spear phishing methods or other social engineering attacks. An organization may need to add new training for staff in specialized cybersecurity roles (PR.AT-02) as the organization adopts AI technologies.

Example #2:

Risk Assessment (ID.RA) AI technologies are used by our adversaries to become faster and more efficient at exploiting a vulnerability from the time it is discovered. Organizations may need to revisit their current processes for how they identify vulnerabilities (ID.RA-01) or for responding to vulnerability disclosures (ID.RA-08).

Example #3:

Asset Management (ID.AM) As AI technologies are adopted and given the new attack surface and ways that AI systems may be exploited, organizations may need to revise their current approach to identifying their data assets (ID.AM-02) or how their data is managed

Consider also where new risks beyond cybersecurity risks might be introduced and how they relate to the NIST Privacy Framework (PF) and the NIST AI Risk Management Framework. Examples #4 and #5 highlight implications across trustworthy AI system characteristics – including privacy - that must be considered before organizations begin to adopt AI for cybersecurity (such as: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair—with harmful bias managed).

Example #4:

Continuous Monitoring (DE.CM) Using AI technology for real time monitoring or anomaly detection may introduce privacy risks to individuals that have not been previously considered. These risks may lead to discrimination, as well as loss of trust, autonomy, or liberty. Other frameworks can be used to more fully contemplate and manage risks beyond cybersecurity, such as identifying the non-monetary costs of trustworthiness issues (MAP 3.2) in the AI RMF and processing data to limit observability and linkability (CT.DP-P1) from the PF.

Example #5:

Identity Management, Authentication, and Access Control (PR.AA) An AI cyber defense system that is learning and making determinations on which incoming emails to block or flag as spam could result in blocking, removing, hiding, or limiting protected speech. AI RMF and PF Subcategories can be used to more fully contemplate and manage privacy risks, such as identifying and understanding unanticipated impacts (MAP 5.2) from the AI RMF and identifying data processing activities that may introduce privacy problems for individuals (ID.RA-P3) from the PF.

2.1. Cyber AI Profile Priorities

For this Profile, NIST is proposing to focus on three sources of AI-related cybersecurity risk that arise from AI technology adoption by organizations including:

1) **Securing AI System Components.** The adoption of AI into existing infrastructures introduces an expanded threat surface, including the vulnerability of AI models themselves, as well as unique and diverse cybersecurity and business challenges. These include the need for new data pipelines, expanded access control and authorization policies, updated employee training, revised service agreements with 3rd party AI providers, and an understanding of new organizational baselines for network activity. The Cyber AI Profile will be available to support organizations in identifying and mitigating the cybersecurity risks associated with AI integration into their organizational ecosystems.
2) **Thwarting AI-enabled Cyber Attacks.** There are a multitude of ways AI is enabling cybersecurity adversaries. For example, AI capabilities are increasing the ease with which adversaries can exploit vulnerabilities as well as expand capabilities for generating new effective attacks, including developing and executing customized attacks that are targeted for a particular organization. The Cyber AI Profile will help organizations focus on activities that build resilience in the face of these new threat vectors.
3) **Using AI for cyber-defense activities.** Security tools are integrating AI-enhanced capabilities to help organizations respond more efficiently to cybersecurity threats and vulnerabilities. Integrating these new capabilities may introduce unanticipated risks, such as a larger, growing infrastructure to monitor, over-reliance on AI detection methods and tools when new, unknown threats are presented, and/or adopting AI tools that are not adapted for an organization's specific needs and data. The Cyber AI Profile will help organizations recognize and assess the efficacy of these new capabilities for cybersecurity and understand the risks to manage when incorporating them into their environment.

2.2. Cyber AI Profile Relationship to Multiple Frameworks

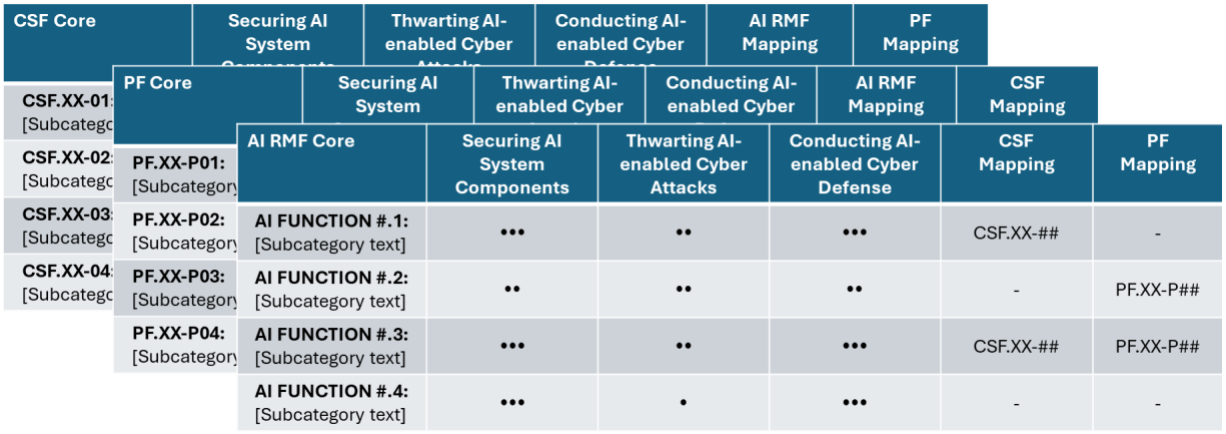
Based on the five CSF Category and Subcategory examples provided earlier, cybersecurity, privacy, and AI practitioners will all be important audiences for the Cyber AI Profile. This indicates a potential need for multi-dimensional views of the Profile content based on the NIST CSF, PF, and AI Risk Management Framework (AI RMF) so that each type of audience can view the Profile through a familiar and relevant lens, identify important areas for collaboration, and also consider how AI risks fit into the broader picture of enterprise risk management and governance. While the initial proposal is for NIST NCCoE to develop the Cyber AI Profile based on the CSF, the relationship to other Frameworks should be considered (Figure 1). Future work on the Cyber AI Profile may also explore workforce implications and the role of the NICE Workforce Framework for Cybersecurity (NICE Framework) and the Privacy Workforce Taxonomy.

Figure 1 - Notional CSF-based Cyber AI Profile (with Mappings to PF and AI RMF)

CSF Core	Securing AI System Components	Thwarting AI-enabled Cyber Attacks	Conducting AI-enabled Cyber Defense	AI RMF Mapping	PF Mapping
CSF.XX-01: [Subcategory text]	•••	••	•••	AI FUNCTION #.#	PF.XX-P##
CSF.XX-02: [Subcategory text]	•	••	••	-	-
CSF.XX-03: [Subcategory text]	••	••	•••	-	PF.XX-P## PF.XX-P##
CSF.XX-04: [Subcategory text]	•••	•••	••	AI FUNCTION #.#	PF.XX-P##

NIST is also exploring ways to highlight the touchpoints between the three frameworks (CSF, PF, AI RMF) to facilitate collaboration where risk areas intersect. One approach is to create three Profiles, each starting with a different framework's core and pointing to relevant aspects of the other two frameworks (Figure 2). Having this multifaceted set of Profiles would provide an organization with a broader perspective of risk management while enabling risk consideration through the lens of their most-familiar framework.

Figure 2 - Notional Suite of Separate Profiles for AI (CSF, PF, and AI RMF)



Another option is to develop the Cyber AI Profile using the CSF and map shared concepts from the PF and AI RMF to relevant CSF Subcategories (Figure 3). This would serve more as a reference to users of the frameworks rather than as an explanatory guide for managing the various dimensions of risk.

Figure 3 - Notional Integrated Cyber AI Profile Oriented by Subcategories with Shared Concepts

CSF Core	PF Core	AI RMF Core	Securing AI System Components	Thwarting AI-enabled Cyber Attacks	Conducting AI-enabled Cyber Defense
CSF.XX-01: [Subcategory text]	**PF.XX-P##:** [Subcategory text]	**AI FUNCTION #.#:** [Subcategory text]	•••	••	•••
CSF.XX-02: [Subcategory text]	-	-	•	••	••
CSF.XX-03: [Subcategory text]	**PF.XX-P##:** [Subcategory text]	-	••	••	•••
CSF.XX-04: [Subcategory text]	**PF.XX-P##:** [Subcategory text]	-	•••	•••	••
-	**PF.XX-P##:** [Subcategory text]	-	•	••	••
-	-	**AI FUNCTION #.#:** [Subcategory text]	•••	•••	•••

Another option that NIST has started exploring would be to create a single, combined Profile for a particular activity common to all three frameworks (Figure 4). This is currently being explored for the domain of data governance and management but could also be an approach to consider risks associated with other shared activities.

Figure 4 - Notional Integrated Activity-Oriented Cyber AI Profile

AI Activity	Securing AI System Components	Thwarting AI-enabled Cyber Attacks	Conducting AI-enabled Cyber Defense
Activity #: [Description]	CSF.XX-##, CSF.XX-##, CSF.XX-##	PF.XX-P## AI FUNCTION #.#	PF.XX-P## AI FUNCTION #.#
Activity #: [Description]	CSF.XX-##	-	CSF.XX-## PF.XX-P##, PF.XX-P## AI FUNCTION #.#
Activity #: [Description]	CSF.XX-##, CSF.XX-##, CSF.XX-##, CSF.XX-##, CSF.XX-##, CSF.XX-## PF.XX-P##	-	-
Activity #: [Description]	-	PF.XX-P##, PF.XX-P##	-
Activity #: [Description]	AI FUNCTION #.#, AI FUNCTION #.#, AI FUNCTION #.#	-	-

And a final consideration would be how to document AI-specific implications for achieving outcomes in the CSF Core as well as Informative References that will help organizations use the Cyber AI Profile (Figure 5).

Figure 5 – Notional AI-specific Implications and Considerations in Cyber AI Profile

CSF Core	Securing AI System Components	Thwarting AI-enabled Cyber Attacks	Conducting AI-enabled Cyber Defense	Informative References / Mappings
CSF.XX-01: [Subcategory text]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[Pointers to related, laws, regulations, guidance, mappings, etc.]
CSF.XX-02: [Subcategory text]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[Pointers to related, laws, regulations, guidance, mappings, etc.]
CSF.XX-03: [Subcategory text]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[AI-specific implications and considerations for achieving this cybersecurity outcome.]	[Pointers to related, laws, regulations, guidance, mappings, etc.]

Of course, the best path forward might involve pursuing multiple options or even a combination.