NIST INFORMATION TECHNOLOGY LABORATORY

SP 800-53 Control Overlays for Securing AI Systems Concept Paper

The advances and potential use cases for adopting artificial intelligence (AI) technologies brings both new opportunities and new cybersecurity risks. While modern AI systems are predominantly software, they introduce different security challenges and risks than traditional software. The security of AI systems is closely intertwined with the security of the IT infrastructure on which they run and operate.

NIST offers a portfolio of guidelines on designing and implementing secure, trustworthy AI, including the AI Risk Management Framework (RMF), guidelines to manage misuse risk from advanced AI (draft), and a taxonomy of AI attacks and mitigations. Feedback received during the Cybersecurity and AI Profile Workshop in April 2025 and ongoing engagement with AI and cybersecurity stakeholders have shown a strong demand for NIST to provide additional implementation-focused guidelines that build on existing resources and frameworks to improve the cybersecurity of AI systems.

To complement ongoing efforts to create a Cybersecurity Framework Profile for AI[1] and support adoption of the AI RMF, NIST is proposing to develop a series of ***Control Overlays for Securing AI Systems (COSAIS)*** using the NIST Special Publication (SP) 800-53 controls.

Why Use SP 800-53 Security Controls and Overlays?

The controls[2] needed to manage risks to AI systems and components will be similar to those required for any type of software. Many organizations are already familiar with [SP 800-53] controls and have institutional processes in place to plan for and assess their implementation. These controls offer both the flexibility to meet unique requirements and a level of specificity that allows for consistent technical implementation.

Control overlays enable organizations and communities of interest to customize the controls (or control baselines) for a specific technology, system, mission space, and environment of operation. Controls can be selected from the [SP 800-53] control catalog, modified to address unique risks or applications, and supplemented to provide application-specific guidance for implementers. The parameter values for assignment and selection operations can also be set.

Using the SP 800-53 controls provides a common technical foundation for identifying cybersecurity outcomes, and developing overlays allows for customization and the prioritization of the most critical controls to consider for AI systems.

[1] As the Cybersecurity Framework Profile for AI effort progresses, NIST will engage with that community of interest to identify how the control overlays may impact the profile's cybersecurity outcomes and vice versa to ensure that both the Profile and overlay provide complementary guidelines for organizations.

[2] Per [OMB Circular A-130], "The safeguards or countermeasures prescribed for an information system or an organization to protect the confidentiality, integrity, and availability of the system and its information or compliance with applicable privacy requirements and manage privacy risks."

Table 1. Relationship between NIST AI and Cybersecurity Efforts

	Control Overlays for Securing AI Systems	Cybersecurity Framework Profile for AI	AI RMF[3]	Managing Misuse Risk for Dual-Use Foundation Models
Purpose	Provide a series of guidelines for implementing SP 800-53 security controls for different uses of AI	Provide a strategic roadmap for organizations to adapt and respond to new or modified cybersecurity risks because of advancements in AI	Provide an approach to establishing and maintaining trustworthiness in AI systems that are used for cybersecurity	Provide guidelines to identify, mitigate, and manage risks of criminal misuse of AI models
Scope	**Security controls** to manage unique risks for users and developers of AI systems in specific scenarios	Priorities and implications for securing AI systems and their components, using AI for cyber defense, and defending against AI-enabled cyber attacks	Identification of risks and opportunities for using AI and suggested actions to mitigate/realize those risks and opportunities	Organizational and model-specific recommendations to manage risks from chemical, biological, radiological, nuclear, and explosive threats (CBRNE), offensive cyber, and other national security-relevant AI capabilities
Audience	Cybersecurity practitioners, AI users, AI developers (dependent on use case)	Cybersecurity senior leadership, CISOs, risk officers, cybersecurity practitioners, AI providers and users	AI actors across the AI lifecycle, including cybersecurity practitioners	AI developers, particularly of frontier models
Assumptions	Organizational/enterprise cybersecurity policy, procedures, and technical controls in place to manage risks	Organizational/enterprise policy, procedures, and technical controls in place to manage cybersecurity risks	Enterprise risk management capabilities are already in place (e.g., cyber, privacy)	Enterprise risk management capabilities are already in place

[3] Includes AI RMF Profiles

The control overlays for securing AI systems will be an implementation-focused series of guidelines to cover different types of AI systems, specific AI system components (e.g., training and test data, model weights, configuration settings) and different use cases. The overlays will focus on protecting the confidentiality, integrity, and availability of information for each use case.[4]

In addition to the SP 800-53 controls, NIST has three foundational AI and cybersecurity publications:

- SP 800-218A, *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile*, serves as a starting point for identifying specific information assets and security practices to translate to controls for AI developers.
- NIST Trustworthy and Responsible AI (AI) 100-2e2025, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, provides a set of AI use cases, attacks, and mitigations.
- Draft NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models, provides organizational and model-specific practices to manage risks of criminal misuse of AI systems.

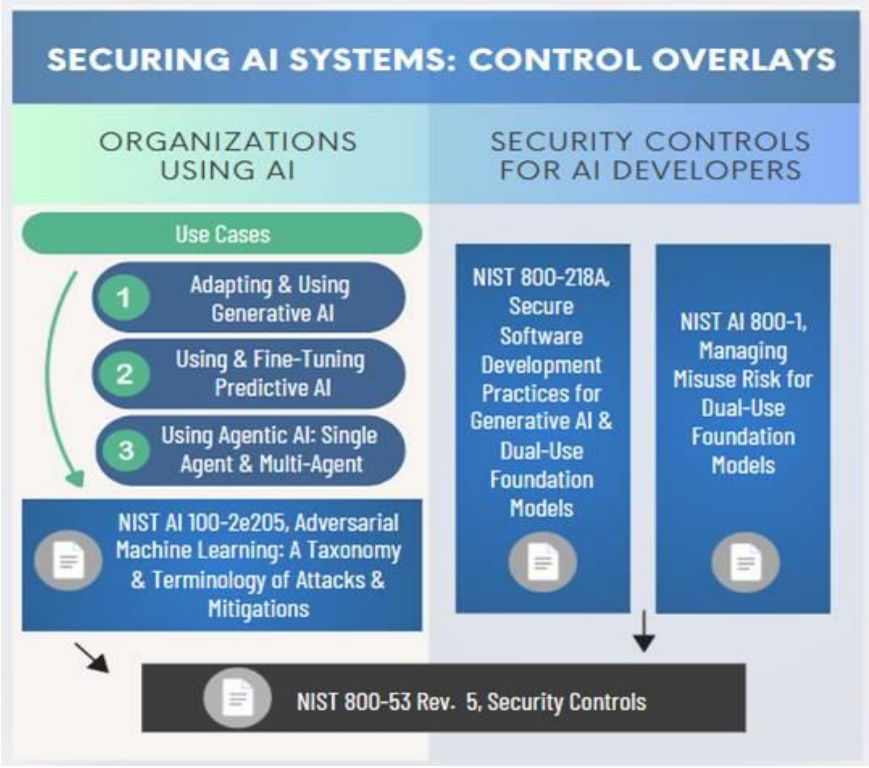

Fig. 1. Relationship between the control overlays and NIST publications

[4] The AI RMF addresses additional risk and trustworthiness considerations (e.g., validity, reliability, safety, accountability, transparency, explainability, interpretability, privacy-enhanced, fair) that are beyond the scope of SP 800-53 security and privacy controls. The series of control overlays can be used in conjunction with the AI RMF for a holistic approach to AI risk management.

As shown in Fig. 1, the control overlays for securing AI systems will leverage AI 100-2e2025, Draft AI 800-1, SP 800-218A, and SP 800-53. The attacks and mitigations from AI 100-2e2025 will inform the selection and tailoring of SP 800-53 controls for four use cases that are designed for organizations using AI. SP 800-218A and Draft AI 800-1 will inform the selection and tailoring of SP 800-53 controls for the use case designed for AI developers.

Proposed AI Use Cases for Security Control Overlays

The intended output is a library of overlays that organizations can use individually or in combination to better manage cybersecurity risks in AI use and development. The overlays will be designed to address specific risks to different types of AI usage in conjunction with an organization's cybersecurity risk management program and existing control implementation. The overlays will not be a comprehensive set of controls for securing an enterprise and will assume that certain controls are already in place (e.g., organization-wide policies, procedures, and implementations of access control for datasets and services, account management, identification and authentication, configuration management, incident response).

NIST proposes an initial set of five use cases for developers and organizations that use AI. Each use case represents a category of common basic scenarios and addresses specific cybersecurity risks. As appropriate, each use case will address the protection of the model, model artifacts, related information assets within the deployment infrastructure, and the security of the output.

Use Case 1: Adapting and Using Generative AI – Assistant/Large Language Model (LLM)	
Audience	Organizations interested in using AI for creating new content (e.g., text, images).
Purpose	Generative AI creates new content (e.g., text, images, audio, video) based on user prompts by learning from large datasets and identifying patterns in the datasets.
Use Case Description	This use case will cover examples of generative AI used for internal business augmentation (e.g., creating summaries, analyzing data) by internal users. A. On-premises hosted LLM with supporting infrastructure: Retrieval-augmented generation (RAG) with proprietary local data sources, mixed local and web sources, and external web sources only. B. Third-party hosted LLM with supporting infrastructure: RAG with proprietary local data sources, mixed local and web sources, and external web sources only.

Use Case 2: Using and Fine-Tuning Predictive AI	
Audience	Organizations using predictive AI systems to analyze historical data to inform decision-making (e.g., for business and service augmentation).
Purpose	Predictive AI uses statistical analytics and machine learning to analyze historical data and predict future outcomes, trends, or behaviors. Examples may include recommendation services, classification services, and business

	workflow efficiency improvements through automated decision-making (e.g., resume review for hiring, credit underwriting).
Use Case Description	These use cases will address cybersecurity risks at three stages of the predictive AI life cycle: (i) model training, (ii) model deployment, and (iii) model maintenance. Each scenario will address a different business workflow example focused on unique cybersecurity risks of using AI systems (e.g., on-premises or third-party hosted AI model, using propriety data or publicly available data). A. Automation of business workflow with predictive AI using a third-party hosted model with proprietary input data: Identify a legitimate need for physical access for authorized users and update the model with refined data for improvement after observation and false positive/false negative rates. B. Automation of business workflow with predictive AI using a third-party hosted model with public-facing input data: Identify "bad actors" and update the model with refined data for improvement after observation and false positive/false negative rates. C. Automation of business workflow with predictive AI using an on-premises model with proprietary data only: Assess the credit worthiness of loan applicants and update the model with refined data for improvement after successful or unsuccessful loan allocations. D. Automation of business workflow with predictive AI using an on-premises model with public-facing input data: Used for hiring purposes (e.g., identifying resumes of candidates to interview/hire based on organization-defined criteria) and updated with refined data for improvement after successful or unsuccessful interview/hires.

Use Case 3: Using AI Agent Systems (AI Agents) – Single Agent	
Audience	Organizations interested in using AI agent systems to automate business tasks and workflows.
Purpose	AI agent systems have the capability for autonomous decision-making and taking action to operate with limited human supervision to achieve complex goals. Characteristics of AI agent systems include the ability to understand context, reason, plan, adapt, and execute tasks.
Use Case Description	This use case will cover examples of AI agent system use. A. Enterprise Copilot - Connected to the user's personal enterprise environment, including emails, files, calendar, or internal enterprise systems (e.g., CRM, IT requests). In addition, Enterprise Copilot can assist the user with common tasks, such as creating calendar events, streamlining workflows, and providing contextual insights. Uses MCP to access data sources. An example of this architecture is available in OWASP Agentic AI Threats and Mitigations. B. Coding Assistant — Understands the enterprise codebase and automates the development of software through natural language commands. Uses MCP to access data sources. This scenario will include:

	a. Connecting to the source code repository to enable the editing of files; b. Interacting with a code repository to create and commit pull requests (PRs), resolve merge conflicts, and similar actions; c. Developing, executing, and fixing unit and integration tests; d. Browsing proprietary and web resources; e. Assisting in the deployment of software.

Use Case 4: Using AI Agent Systems (AI Agents) – Multi-Agent	
This use case is currently the least mature in terms of adoption, but it is expected to evolve and be refined as understanding deepens, and implementation challenges are identified and addressed.	
Audience	Organizations interested in using AI agent systems to automate complex business tasks and workflows with multiple interacting workstreams.
Purpose	Multi-agent AI systems have the capability for autonomous decision-making and have multiple agents working in concert taking action to operate cooperatively with limited human supervision to achieve complex goals. Characteristics of multi-agent AI systems include the ability to understand context, reason, plan, adapt, coordinate actions, and execute tasks.
Use Case Description	Multiple agents working in concert to extract data and take actions on the basis of that data, leveraging standardized protocols such as MCP or A2A for agent-to agent communication. For example, based on reference architectures from OWASP Multi-Agentic System Threat Modeling Guide: Robotic Process Automation Expense Reimbursement Agent. Multiple agents responsible for extracting information from expense claims (e.g., submitted receipts, forms), validating the claims against company policy leveraging RAG, and routing approved claims for payment to automate the employee reimbursement workflow in the enterprise.

Use Case 5: Security Controls for AI Developers	
Audience	AI developers
Purpose	NIST developed an SSDF Community Profile: Secure Software Development Practices for Generative AI and Dual-Use Foundation Models (SP 800-218A), which identifies critical-for-security model artifacts and good practices for securing them. NIST CAISI published Managing Misuse Risk for Dual-Use Foundation Models (Draft AI 800-1) as a resource for AI developers. A mapping from the security controls in SP 800-53 to these artifacts and practices in SP 800-218A and in Draft AI 800-1, resources pending, can benefit the community of AI developers and allow for effective risk management built upon existing organizational practices.

NIST Overlays Securing AI Slack Channel

The Security Control Overlays for AI Systems project will leverage a newly launched NIST Overlays for Securing AI Slack channel (#nist-overlays-securing-ai), which is a hub for cybersecurity and AI communities to discuss the development of these overlays.

Join the NIST AI Overlay Slack Channel to get updates, engage in facilitated discussions with the NIST principal investigators and other subgroup members, share ideas, provide real-time feedback, and contribute to overlay development. All interested parties are welcome.

Next Steps and Engage With NIST

NIST requests your feedback on the proposed high-level use cases and potential additional future work. Specifically, NIST seeks feedback on the following:

- How well do the use cases capture representative types of AI adoption for user communities, what potential gap areas need to be addressed?
- To what extent do the architectures and example use cases reflect real-world adoption patterns, and where might there be gaps or issues?
- How should NIST prioritize overlay development for the use cases?
- Are there additional areas or use cases to consider for future work?

Feedback to the questions above and on the concept paper should be sent to overlays-securing-ai@list.nist.gov, and can also be shared in the Slack channel. Learn more about the Control Overlays for AI Project, Slack Space, and how to engage at https://csrc.nist.gov/preview/projects/cosais.

Based on feedback, NIST will start the first use case with the goal of issuing a public draft for comment in early FY26, one use case at a time. A public workshop will also be held during the public comment period for ongoing stakeholder engagement.