Artificial Intelligence (AI) Risk Management Framework (RMF)
National Institute of Standards and Technology (NIST)
Request for Information (RFI)

Cybersecurity Innovation Program (CIP)
United States Department of Veterans Affairs (VA)
April 29, 2022

Request for Information Supporting Points

Question 1. Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases.

The AI RMF successfully covers a broad spectrum of risks and use cases. Nonetheless, VA CIP recommends considering additional information to illustrate the multifaceted challenge of securing AI systems. Some examples include the following:

- VA CIP recommends that NIST continuously update the AI RMF taxonomy to align with emerging AI research and understanding, as the relationship between interconnected AI characteristics become clearer. Namely, as the relationship between security, safety, and privacy becomes better defined, NIST may need to update or clarify these characteristics and provide additional guidance on specific technical features to promote risk management across Technical, Socio-Technical, and Trustworthiness characteristics.
- In previous publications, other organizations have noted the risk of organizational overreliance on third-party vendors for AI capabilities. VA CIP similarly notes that this overreliance may lead to concentration risk, system risk, and supply chain risk. Such elements may lead to increased opacity and less control over how AI systems function. Additionally, supply chain security vulnerabilities from third-party software and hardware should be critical considerations in a complete risk management framework. NIST should factor these and other similar operational concerns into an existing risk category (e.g., Accountability, Transparency, Resilience, or Machine Learning [ML] Security) as the topic deserves additional emphasis. Following this recommendation, VA CIP suggests highlighting the latest guidance from the NIST AI Secure Software Supply Chain and Secure Software Development Framework (SSDF), which includes leading practices like requiring a Software Bill of Materials (SBOM).
- The Resilience or ML Security section states that "ML security concerns relate to the exfiltration of models, training data, or other intellectual property through AI system endpoints." We recommend mentioning that these risks are present throughout the whole AI model development lifecycle and may not rely on model endpoints alone (e.g., backdoor attacks during the model development phase or data poisoning attacks pre-deployment). In addition to data and model exfiltration, NIST should also mention that resilience includes preventing the tampering of datasets and securely processing, storing, and sharing the data that fuels AI/ML.
- The Privacy section currently cites privacy as related to "the processing of data." VA CIP recommends expanding this definition, as privacy risks can occur due to other vulnerabilities beyond capturing, transmitting, and storing data. For example, an inference attack targets an AI model to reveal

underlying sensitive data. As such, privacy concerns can come from both the model and data, and organizations should be vigilant of both.

- Although NIST recommends using other previously developed guidance, the framework should provide more information about how cybersecurity and privacy threats may apply to an AI system. The current draft focuses heavily on the resilience and reliability of a model but does not provide a clear explanation of the multiple areas that can impact security. NIST should provide additional context that security concerns derive from aspects like securing hardware, employing proper software security hygiene, securing data at rest, ensuring strong encryption protocols, using data sharing mechanisms that preserve confidentiality, and considering other traditional cybersecurity challenges that still apply to AI.

Question 2. Whether the AI RMF is flexible enough to serve as a continuing resource considering evolving technology and standards landscape.

The AI RMF is a useful and flexible resource. VA CIP emphasizes the need to keep the document relevant by continuously integrating ideas and guidance across several domains through a periodic maintenance process:

- VA CIP recommends continuously aligning with recent legal and policy regulations at the Federal, state, and local levels (e.g., Int. 1894-2020A, which requires algorithms used for hiring purposes to undergo a bias audit in New York City).[1] Also, the framework should monitor industry-specific guidance (e.g., Health Insurance Portability and Accountability Act [HIPAA] for privacy and deidentification standards, Cancer Imaging Archives standards for Digital Imaging and Communications in Medicine [DICOM] images, and applicable regulations and standards from the financial sector). While a discussion of these specific standards is out of scope for the AI RMF, the granularity of standards for different subdisciplines of AI should be referenced in the Practice Guide.
- VA CIP recommends continuing to track evolving guidance for related fields that overlap with AI (e.g., the NIST Cybersecurity and Privacy frameworks, Zero Trust Architecture).
- VA CIP recommends continuing to track evolving global guidance from global organizations and government agencies (e.g., the United Kingdom’s National Cyber Security Centre [NCSC]).

Question 3. Whether the AI RMF enables decisions about how an organization can increase the understanding of, communication about, and efforts to manage AI risks.

The AI RMF accomplishes this objective through the depictions of key stakeholder groups, the AI risk taxonomy, and Functions. We have the following recommendations for enhancing the AI RMF for this purpose:

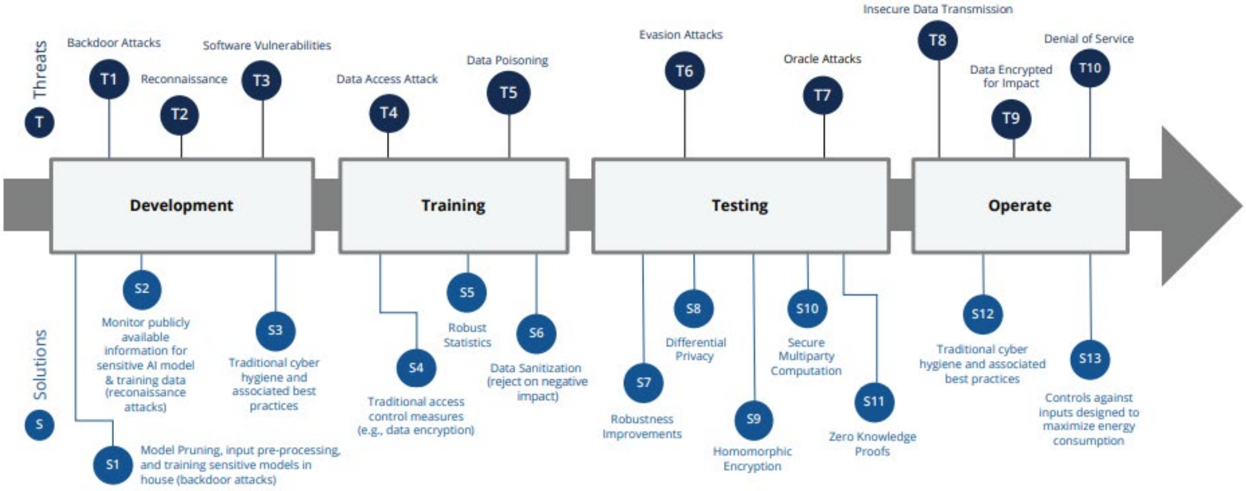
- ID 1 for the Measure Function is centered around the identification of relevant risk metrics. The second subcategory elaborates on this, specifically mentioning “… including technical measures of performance…”. It may be valuable to also emphasize that recent research is identifying methods for quantifying technical risks such as AI security, as well as non-technical principles such as the socio-technical principle of privacy (e.g., k-anonymity).[2] These are comparatively less well-known and can be useful for organizations to leverage in parallel to traditional performance metrics.
- We recommend providing a high-level overview of novel AI attack techniques (e.g., oracle attacks, backdoor attacks, model extraction attacks) threat actors may use across AI lifecycle phases (e.g.,

[1] James Anelli et al. 2022 “New Laws Impacting Hiring and Promoting in New York City,” White and Williams LLP, https://www.jdsupra.com/legalnews/new-laws-impacting-hiring-and-promoting-9762083/

[2] Hatamizadeh, A. et al. 2022. Do Gradient Inversion Attacks Make Federated Learning Unsafe? https://arxiv.org/pdf/2202.06924.pdf

develop, train, test, operate), including potential mitigation techniques (i.e., solutions) and the risks associated with these attacks. For more information see Figure 1: AI Threats and Solutions.

Figure 1: AI Threats and Solutions


Threats
T
Backdoor Attacks
T1
Reconnaissance
T2
Software Vulnerabilities
T3
Data Access Attack
T4
Data Poisoning
T5
Evasion Attacks
T6
Oracle Attacks
T7
Insecure Data Transmission
T8
Data Encrypted for Impact
T9
Denial of Service
T10
Development
Training
Testing
Operate
Solutions
S
S1
Model Pruning, input pre-processing, and training sensitive models in house (backdoor attacks)
S2
Monitor publicly available information for sensitive AI model & training data (reconaissance attacks)
S3
Traditional cyber hygiene and associated best practices
S4
Traditional access control measures (e.g., data encryption)
S5
Robust Statistics
S6
Data Sanitization (reject on negative impact)
S7
Robustness Improvements
S8
Differential Privacy
S9
Homomorphic Encryption
S10
Secure Multiparty Computation
S11
Zero Knowledge Proofs
S12
Traditional cyber hygiene and associated best practices
S13
Controls against inputs designed to maximize energy consumption

- We recommend the discussion of interdisciplinary efforts and roles within an organization in either the AI RMF itself or one of its companion documents (e.g., the Practice Guide). Additionally, security best practices should be understood by all team members that develop, train, test, and operate AI models: see Table 1 for more information on these roles. It should be noted that 'Roles' and 'Descriptions' for these profile teams are solely intended as symbolic representation for the distinctions between teams that develop security infrastructure, machine-learning development infrastructure, and respective end-users (data scientists) within an AI development ecosystem.
- We recommend adding that organizations should create maintain a detailed inventory of the AI / ML applications and systems developed, trained, and deployed within an organization. Additionally, we recommend that the inventory should contain a log of cybersecurity and risk-associated characteristics (e.g., what controls are in place to protect against and detect data poisoning attacks?), including questions contained within the General Services Administration (GSA's) Algorithmic Impact Assessment (AIA) tool (e.g., Is there a system in place to ensure the secure transfer of data across multiple networks?).

Table 1: Description of Profile Teams

Profile Teams	Roles	Description	Project Tasks
Secure Software Development	• Frontend developers • Backend developers • Data engineers • Cloud architects • Cybersecurity engineers	Designed for traditional software development. Develops logic-based software that does not require sophisticated AI components.	• Establishing secure development pipelines • Assessing and validation security of new tools • Building backend components • Cyber hygiene policy implementation • Access control list management
Machine Learning Operations (MLDevOps)	• Data scientists • Data engineers • Cloud architects • Data labelers	Intended to support large scale AI enabled software projects such as computer vision, predictive healthcare modeling, etc.	• Curating ML datasets for training • Building data pipelines for applications • Building ML models • Managing and monitoring model deployment
Operational Data Science Research Team	• Data scientists • Data engineers • Cloud architects	Focused on data specific product generation. Intended to be imbedded within operations working directly for an end-user	• Data science process • Building data pipelines for applications • Building ML models • Building and deploying reports, basic web applications, dash boards

Question 4. Whether the functions, categories, and subcategories are complete, appropriate, and clearly stated.

Overall, the functions, categories, and subcategories are well considered and clearly stated. Our recommendations below address the "Map," "Measure," and "Govern," function:

- *Map Function:* Security threats are increasing both in complexity and frequency and target different stages of AI system development to achieve their goals. The inclusion of guidance that identifies and groups risk types associated with different AI lifecycle phases may be useful, as the impact and management of these risks may vary greatly by phase (see Figure 1: AI Threats and Solutions). In addition, we recommend the inclusion of cost categories related to AI security and privacy that may be overlooked at the start of an AI project, such as continued security surveillance scanning, automating data privacy measures at scale for model retraining, etc.

- *Measure Function:* While AI risk attributes such as transparency and accuracy may be incrementally addressed and remedied over time, certain risks such as security or privacy must be prioritized from the beginning due to their potential impact. For example, the exfiltration of a training dataset containing the private information of individuals cannot reasonably be remedied. As such, these impacts should be weighed heavily.

- *Govern:* We recommend strongly emphasizing the need for not only organizational accountability, but also system accountability. This includes a well-documented audit trail that captures all decisions made in the design in an AI system. From CIP's security standpoint, this may also extend beyond the AI system itself, but also encompass the environment it is developed and deployed in (e.g., the configuration of the host cloud infrastructure or an organization's on-premises systems)

Question 5. Whether the AI RMF is in alignment with or leverages other frameworks and standards such as those developed or being developed by IEEE or ISO/IEC SC42.

- This recommendation follows from our comment made in response to Question #2. ISO/IEC SC42 mentions legal and regulatory guidance as considerations when establishing the external context of an organization. The NIST AI RMF functions may benefit from also capturing this. We recommend including language around reviewing the legal and regulatory landscape in the Map function.
- The Measure function in the NIST AI RMF mentions the need to regularly assess and update metrics for appropriateness. ISO/IEC SC42 also notes that measurement methods are constantly evolving and should be evaluated on a regular basis. We recommend noting that organizations should evaluate the advantages and disadvantages of relying on certain metrics and adding or modifying them as necessary.
- ISO/IEC SC42 A.9 identifies the facets of accountability as organizational and system accountability. Similar to our comment in Question #4, we recommend adding language to account for system accountability.
- ISO/IEC SC42 A.5, lists several standards relating to "Privacy", "Security," and development of a "Privacy Assessment document," which can be used to develop more comprehensive guidance for Sections 5.2.3 - "Privacy," and Section 5.2.4 – "Safety," for the NIST AI RMF.

Question 6. Whether the AI RMF is in alignment with existing practices, and broader risk management practices.

- From CIP's initial review, the NIST AI RMF draft is currently aligned to existing guidance / risk management practices, including but not limited to, General Services Administration (GSA) Algorithmic Impact Assessment (AIA) Questionnaire, Microsoft's Best Practices for AI Security Risk Management, Health and Human Services (HHS) AI Playbook, and the European Telecommunications Standards Institute (ETSI) Secure AI guidance.
- While the NIST RMF is a comprehensive draft and is not intended to be an exhaustive checklist, it may be helpful to reference additional materials that guide developers through each step of the lifecycle phase relating to privacy and security perspectives and best practices (e.g., appropriate data hygiene, routine testing, and updating after deployment) to ensure that AI systems are safely being deployed.
    - Per this recommendation, we cite examples of practices in-development at VA-CIP for secure AI development in "Question 8," of this response.

Question 7. What might be missing from the AI RMF.

- Please refer to "Question 8".

Question 8. Whether the soon to be published draft companion document citing AI risk management practices is useful as a complementary resource and what practices or standards should be added.

- A companion document to the AI RMF would be invaluable to AI stakeholders. This document would provide practical guidance and templates for operationalizing high-level theoretical guidance to further organizational missions.
- There are several types of guidance that may be useful to include in a Practice Guide. From VA CIP's perspective, we have had success within the internal VA AI community in providing both process-based and tool-based guidance. Our process-based guidance includes checklists targeted towards data science teams that document the steps necessary for developing secure AI systems, as illustrated in *Figure 2*. These consolidate both leading industry and internal guidance.

- ○ Secure AI Development Checklist (Figure 2);
- ○ Secure AI Development Tools Guidance (Figure 3); and
- ○ Secure AI Data and De-identification Technique Guidance (Figure 4)

Figure 2: Example of Secure AI Development Checklist (for illustration purposes only)

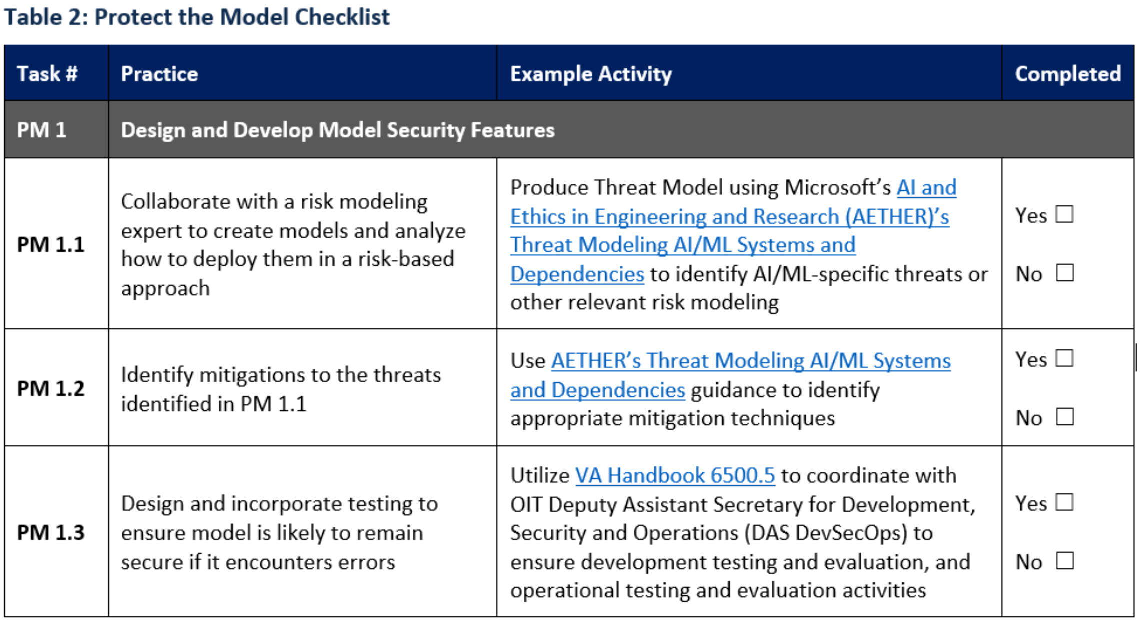


Table 2: Protect the Model Checklist

Task #	Practice	Example Activity	Completed
PM 1	**Design and Develop Model Security Features**		
PM 1.1	Collaborate with a risk modeling expert to create models and analyze how to deploy them in a risk-based approach	Produce Threat Model using Microsoft's AI and Ethics in Engineering and Research (AETHER)'s Threat Modeling AI/ML Systems and Dependencies to identify AI/ML-specific threats or other relevant risk modeling	Yes ☐ No ☐
PM 1.2	Identify mitigations to the threats identified in PM 1.1	Use AETHER's Threat Modeling AI/ML Systems and Dependencies guidance to identify appropriate mitigation techniques	Yes ☐ No ☐
PM 1.3	Design and incorporate testing to ensure model is likely to remain secure if it encounters errors	Utilize VA Handbook 6500.5 to coordinate with OIT Deputy Assistant Secretary for Development, Security and Operations (DAS DevSecOps) to ensure development testing and evaluation, and operational testing and evaluation activities	Yes ☐ No ☐

Our tools-based guidance is centered around specific topics such as secure AI development. These synthesize current regulations and research and highlight useful references, industry and open-source tools, and resources for operationalizing them. *Figure 3* provides a brief snapshot of what these include.

Figure 3: Example Secure AI Development Tools Guidance (for illustration purposes only)

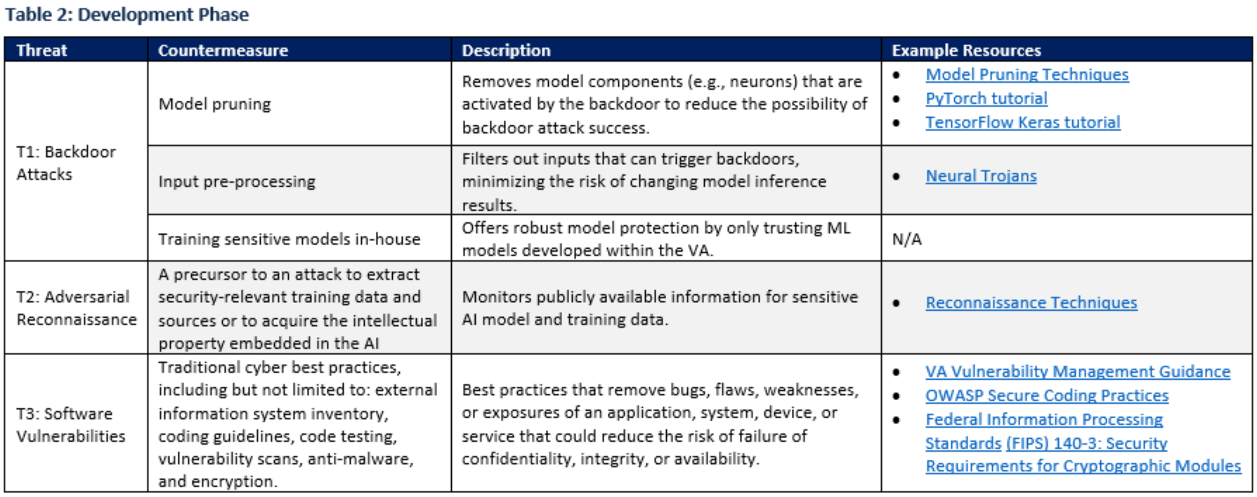


Table 2: Development Phase

Threat	Countermeasure	Description	Example Resources
T1: Backdoor Attacks	Model pruning	Removes model components (e.g., neurons) that are activated by the backdoor to reduce the possibility of backdoor attack success.	• Model Pruning Techniques • PyTorch tutorial • TensorFlow Keras tutorial
	Input pre-processing	Filters out inputs that can trigger backdoors, minimizing the risk of changing model inference results.	• Neural Trojans
	Training sensitive models in-house	Offers robust model protection by only trusting ML models developed within the VA.	N/A
T2: Adversarial Reconnaissance	A precursor to an attack to extract security-relevant training data and sources or to acquire the intellectual property embedded in the AI	Monitors publicly available information for sensitive AI model and training data.	• Reconnaissance Techniques
T3: Software Vulnerabilities	Traditional cyber best practices, including but not limited to: external information system inventory, coding guidelines, code testing, vulnerability scans, anti-malware, and encryption.	Best practices that remove bugs, flaws, weaknesses, or exposures of an application, system, device, or service that could reduce the risk of failure of confidentiality, integrity, or availability.	• VA Vulnerability Management Guidance • OWASP Secure Coding Practices • Federal Information Processing Standards (FIPS) 140-3: Security Requirements for Cryptographic Modules

Our recommendations also consider the link between data and privacy. It is common for datasets to contain sensitive and regulated information in health care and other use cases. There are several regulatory and theoretical approaches to removing identifiers and anonymizing individuals in a dataset. From the perspective of HIPAA, Section 164.514(a) of the HIPAA Privacy Rule[3] provides a standard for de-identification of protected health information. In *Figure 4*, we demonstrate how organizations can survey de-identification solutions in line with HIPAA Privacy Rule, which mask, redact, and blur personally identifiable characteristics within data sets. Solutions are then evaluated to identify appropriateness of primary de-identification techniques for an organization, input datatypes, strengths, and limitations. Solutions can also be sourced through exhibited past performance at similar enterprises in the federal or private sector.

Figure 4. Example Secure AI Data and De-identification Technique Evaluation

Solution Tool/Vendor	Primary De-Identification Technique	Datatypes	Strengths	Considerations
Vendor 01	• Masking • Blurring	• Unstructured text • Image / Video	• Flexible computational power to meet input demand • Configurable for specific use cases • Integrated cybersecurity services (e.g., encryption)	• Data must travel to the cloud • High level of expertise / configuration required
Vendor 02	• Masking	• Unstructured text • Image / Video	• +65 file formats (e.g., xml, csv, pdf) • Low code • Integrated cybersecurity services (e.g., encryption)	• Data must travel to the cloud • Non-health care specific solution
Vendor 03	• Differential privacy	• Tabular data	• Minimal end-user intervention within model parameters • Deployable on any off-the-shelf optimizer	• High level of expertise required • Not readily implemented into understandable package/GUI
Vendor 04	• Suppression	• Tabular data	• Global and local transformation options • Intuitive cross-platform graphical tool	• High-dimensional analysis but, supports limited types of files
Vendor 05	• Redaction	• Image (e.g., DICOM)	• Low code • International userbase for collaboration	• Requires data to be uploaded to third-party environment
Vendor 06	• Masking	• Unstructured text	• Masking allows for replacement of PHI with realistic surrogates • Highly configurable rule-based approach	• Accuracy for misspelled PHI and location identifiers may be difficult to detect

Question 9. Others?

No additional comment.

[3] "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule" https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#_ednref4