Meeting M-24-10 Deadlines: Can we adapt FIPS Pub 199 to Classify AI Risk?

Office of Management and Budget (OMB) issued M-24-10 Memorandum for the Heads of Executive Departments and Agencies for advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence (AI). By December 1, 2024, agencies must implement the minimum practices for safety-impacting and rights-impacting AI, or else stop using any AI in their operations that is not compliant with the minimum practices.

One of the critical questions is: "How should a Government Agency provide an authority to operate for AI systems by December 1, 2024?"

Using National Institute of Standards and Technology (NIST) standards such as the NIST AI Risk Management Framework (RMF) is the right place to begin. However, the AI RMF provides a high-level roadmap to think about AI risk management but does not offer prescriptive guidance and specific actions that must be taken by a Chief AI Officer (CAIO), Chief Digital Officer (CDO), Chief Information Officer (CIO) or Chief Information Security Officer (CISO).

Federal agencies have a lot of experience with cybersecurity risk management with a mature set of policies, procedures, and guidance on how to assess and accredit Information Technology (IT) systems. Could that experience not be leveraged to managed risks associated with emerging AI technologies?

- What if, we could connect NIST AI RMF with NIST RMF & NIST SP 800-53, with extensions and tailoring to address AI-specific risks like Safety, Bias, and Explainability (SBE)?

- What if, agencies could meet the M-24-10 mandated deadlines by tailoring and augmenting existing standards like FIPS Pub 199 that define risks from loss of Confidentiality, Integrity, and Availability (CIA), by adding a new AI-specific triad: Safety, Bias and Explainability (SBE)?

- Can agencies then start to rapidly execute the needed RMF steps: Prepare, Categorize, Select, Implement, Assess, Authorize, and finally monitor?

The cloud, security, and compliance experts at stackArmor, Inc. have developed the ATO for AI™ governance model and the first AI Risk Management Center of Excellence (CoE) to help agencies deploy safe and secure AI systems.

The figure below provides an overview of stackArmor's Authorization To Operation (ATO) for AI governance model that begins with Trustworthy AI and uses the AI RMF risk categories, mapped to NIST SP 800-53 controls, with overlays and tailoring to accommodate the AI-specific risk triad.

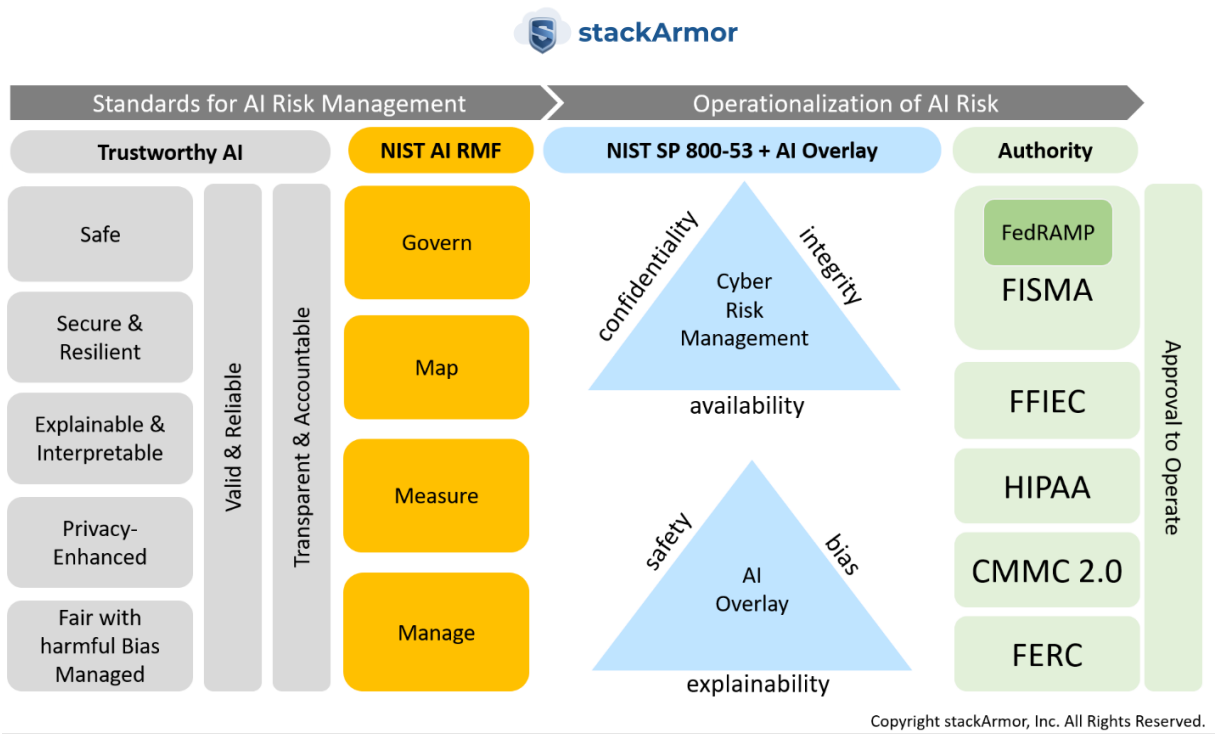

Figure 1: ATO for AI™ Governance Model to Help Federal Agencies meet M-24-10 Requirements for AI Risk Management Plans and Approaches

This blog explains how this mapping might work in practice. The NIST AI RMF's Map function is designed to help set the context and risks related to the context are identified. By connecting NIST AI RMF to NIST Risk Management Framework (RMF), we can rapidly start performing risk management actions using familiar tools, which however need to be tailored and augmented to deal with AI-specific risks. The NIST AI RMF **Map** function can be realized using the NIST RMF **Categorize** step to inform organizational risk management processes and tasks by determining the adverse impact. **FIPS Pub 199** is used by Federal agencies as a standard to perform the security categorization for information and information systems across the government as well as commercial organizations providing cloud computing services to Federal agencies.

- What if, using the groundwork laid out by FIPS Pub 199, the CIA triad was further enhanced by the AI-specific SBE triad?

The combination of the CIA triad and the SBE triad provides Agencies with a familiar methodology to rapidly enable security and risk practitioners to start implementing risk management plans and procedures. This rapid adoption of modified, but familiar methodologies, would allow agencies to meet the ambitious goals and timelines of M-24-10. The combination of CIA and SBE ensures that agencies are compliant with the seven Trustworthy AI pillars as outlined in the NIST AI RMF.

Adding the SBE Triad to FIPS 199

It is important to provide Federal Agencies and the Chief AI Officer (CAIO) with the tools necessary to meet the mandated timelines of M-24-10. To inform the path forward towards Risk Assessment of AI systems, the following is an example of FIPS 199, modified with the proposed AI SBE Triad.

	Potential Impact		
Objective	**Low**	**Moderate**	**High**

Confidentiality Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information. [44 U.S.C., SEC. 3542]	The unauthorized disclosure of information could be expected to have a limited adverse effect on organizational operations, organizational assets, or individuals.	The unauthorized disclosure of information could be expected to have a serious adverse effect on organizational operations, organizational assets, or individuals.	The unauthorized disclosure of information could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets, or individuals.
Integrity Guarding against improper information modification or destruction, and includes ensuring information nonrepudiation and authenticity. [44 U.S.C., SEC. 3542]	The unauthorized modification or destruction of information could be expected to have a limited adverse effect on organizational operations, organizational assets, or individuals.	The unauthorized modification or destruction of information could be expected to have a serious adverse effect on organizational operations, organizational assets, or individuals.	The unauthorized modification or destruction of information could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets, or individuals.
Availability Ensuring timely and reliable access to and use of information. [44 U.S.C., SEC. 3542]	The disruption of access to or use of information or an information system could be expected to have a limited adverse effect on organizational operations, organizational assets, or individuals.	The disruption of access to or use of information or an information system could be expected to have a serious adverse effect on organizational operations, organizational assets, or individuals.	The disruption of access to or use of information or an information system could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets, or individuals.
Safety The term "safety-impacting AI" refers to AI whose output produces an action or serves as a principal basis for a decision that has the potential to significantly impact the safety of: 1. Human life or well-being, including loss of life, serious injury, bodily harm, biological or chemical harms, occupational hazards, harassment or abuse, or mental health, including both individual and community aspects of these harms; 2. Climate or environment, including irreversible or significant environmental damage; 3. Critical infrastructure, including the critical infrastructure sectors defined in Presidential Policy Directive 2159 or any successor directive and the infrastructure for voting and protecting the integrity of elections; or, 4. Strategic assets or resources, including high-value property and information marked as	The output produced from AI could be expected to have a limited safety implications on organizational operations, organizational assets, or individuals.	The output produced from AI could be expected to have serious safety implications on organizational operations, organizational assets, or individuals.	The output produced from AI could be expected to have catastrophic safety implications on organizational operations, organizational assets, or individuals.

sensitive or classified by the Federal Government. [OMB Memo M-24-10]			
Bias The term "rights-impacting AI" refers to AI whose output serves as a principal basis for a decision or action concerning a specific individual or entity that has a legal, material, binding, or similarly significant effect on that individual's or entity's: 1. Civil rights, civil liberties, or privacy, including but not limited to freedom of speech, voting, human autonomy, and protections from discrimination, excessive punishment, and unlawful surveillance. 2. Equal opportunities, including equitable access to education, housing, insurance, credit, employment, and other programs where civil rights and equal opportunity protections apply; or 3. Access to or the ability to apply for critical government resources or services, including healthcare, financial services, public housing, social services, transportation, and essential goods and services. [OMB Memo M-24-10]	The output produced from AI could be expected to have a limited rights impacting effects on organizational operations, organizational assets, or individuals.	The output produced from AI could be expected to have a serious rights impacting effects on organizational operations, organizational assets, or individuals.	The output produced from AI could be expected to have a catastrophic or severe rights impacting effects on organizational operations, organizational assets, or individuals.
Explainability AI systems should be explainable, and adhere to the following four principles: Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes. Meaningful: A system provides explanations that are understandable to the intended consumer(s). Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process. Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output. [NISTIR 8312]	The lack of explainability of the output of the AI could be expected to have a limited adverse effect on organizational operations, organizational assets, or individuals.	The lack of explainability of the output of the AI could be expected to have a serious adverse effect on organizational operations, organizational assets, or individuals.	The lack of explainability of the output of the AI could be expected to have a catastrophic or severe adverse effect on organizational operations, organizational assets, or individuals.

Figure 2: Potential Impact Definitions for Security and Safety Objectives for AI

Risk Categorization Applied to an AI System

Example: An Agency procurement department is utilizing an AI Engine to facilitate the development and review of pre-solicitation phase contract information against released RFPs and capture routine administrative information. The contract and administration information has the potential to contain

sensitive information including (PII, trade secrets, financial information, contractor proposal information).

The management team within the contracting organization determines that:

i. For the sensitive contract information, the potential impact from a loss of:
   a. Confidentiality is moderate
   b. Integrity is moderate
   c. Availability is low

ii. For the routine administrative information (non-privacy-related), the potential impact from loss of:
   a. Confidentiality is low
   b. Integrity is low
   c. Availability is low.

The resulting CIA security categories, of these information types are expressed as:

SC contract information = {(confidentiality, MODERATE), (integrity, MODERATE), (availability, LOW)}, and
SC administrative information = {(confidentiality, LOW), (integrity, LOW), (availability, LOW)}.

When the management team reviews the contracting AI Engine looking at the SBE Triad they determine that:

i. For the sensitive contract information, the potential impact on:
   a. Safety is low: The stage of the procurement process the AI is injected into poses a low risk that its outputs could cause harm to human life, then environment, or strategic assets.
   b. Bias is high: The impact of potential bias from the AI's outputs could directly and severely impact the equal opportunity and ability to compete for government contracts.
   c. Explainability is moderate: The lack of explainable results of the AI's outputs could have a serious impact on the procurement teams' ability to understand or even see the results for proposals that are disqualified.

ii. For the routine administrative information (non-privacy-related), the potential impact on:
   d. Safety is low: The stage of the procurement process the AI is injected into poses a low risk that its outputs could cause harm to human life, then environment, or strategic assets.
   e. Bias is moderate: The AI system has the potential to inject bias into the process based on exclusion of contractors based on factors captured within the proposals.
   f. Explainability is low: The lack of explainable results of the AI's outputs could have a minimal impact on the operations of the AI.

The resulting SBE security categories, of these information types are expressed as:

SC contract information = {(Safety, LOW), (Bias, HIGH), (Explainability, MODERATE)}, and
SC administrative information = {(Safety, LOW), (Bias, MODERATE), (Explainability, LOW)}.

The categorization information to visualize FIPS 199 with both the CIA and SBE Triads in use, is summarized below.

Risk Categorization						
Information Type	**Confidentiality**	**Integrity**	**Availability**	**Safety**	**Bias/Rights**	**Explainability**
Contract information	Moderate	Moderate	Low	Low	High	Moderate
Administrative information	Low	Low	Low	Low	Moderate	Low
Risk Categorization of System	Moderate	Moderate	Low	Low	High	Moderate

Figure 3: Summary of AI Risk Categorization as Part of AI Risk Assessment

As you can see in the above example the combination of risk categorization using both CIA and SBE Triads can be an extremely valuable resource for Agencies when reviewing the potential impacts of leveraging AI services. The agency now has the tools available to decide to either categorize the overall system as a Moderate (based on the CIA triad) or a High (based on the SBE triad).

Summary

With the looming deadlines of M-24-10, Agencies must rapidly implement easily understood measures that can be quickly and efficiently rolled out to meet the presidential mandates. Exploring ways to extend, augment and tailor existing risk management processes and frameworks will accelerate the process of implementing risk assessment and authorization activities. The information provided in this artifact is to spur discussion and make collective progress towards ensuring the safe and secure induction of AI systems into production.

If you a CFO Act Agency interested in learning more how you can rapidly assess AI risk, please contact us to schedule a free briefing or download our ATO for AI™ white paper.

About stackArmor

stackArmor has over ten years of security and compliance engineering experience in helping public sector organizations accelerate their ability to meeting FedRAMP, FISMA/RMF, DOD Cloud Computing SRG, and CMMC requirements. At the heart of stackArmor's approach is the operationalization of NIST SP 800-53 security controls through a ATO Factory pattern that uses standardization, automation, and specialized teams of experts to reduce the time and cost of ATO projects by 40%. Visit https://www.stackArmor.com to learn more.

ATO for AI™ in the News

FedScoop	"Biden issued his historic EO on artificial intelligence. Now comes the hard part, experts say"
Cyber Defense Magazine	"Navigating Secure Adoption of AI Across Government and Connected Infrastructure"
Nextgov/FCW	"The federal government is already using AI — it's time for a formal process to ensure the technology is safe"

Updated on 4/16/2024 with minor changes for improved readability.