credo | ai

Feb 2, 2024

National Institute of Standards and Technology
Department of Commerce
100 Bureau Drive, Mail Stop 8970
Gaithersburg, MD 20899-8970

RE: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

About Credo AI

Founded in 2020, Credo AI's mission is to empower enterprises to responsibly build, adopt, procure and use AI at scale. Credo AI's pioneering software AI governance platform helps enterprises from Global 2000s to small- and medium- sized enterprises measure, monitor and manage AI risks, while ensuring compliance with emerging global regulations and standards, such as the NIST AI Risk Management Framework (AI RMF), the European Union (EU) AI Act, and the work of international standard setting bodies such as the ISO and IEEE.

We are encouraged by the evolution of AI governance and guardrails discussions over the past 18 months. Our mission at Credo AI aligns closely with the goals of the White House Executive Order and NIST's mandate to establish guidelines and best practices to promote consensus-based industry standards in the development and deployment of safe, secure, and trustworthy AI systems.

Credo AI is proud to have partnered closely with NIST by both providing feedback on NIST's AI RMF 1.0 and operationalizing it in our product. Governance tools like ours ensure that companies of all sizes, across industries, and in every stage of Responsible AI design, can adopt it. We look forward to our continued cooperation with NIST, providing industry insights to help develop standards that satisfy the needs of the community and advance responsible AI adoption globally.

Executive Summary

The directive E.O. 14110 mandates NIST to establish guidelines and best practices for safe, secure, and trustworthy AI systems, particularly focusing on generative AI. This document draws from Credo AI's experience working with enterprises to operationalize responsible AI at scale, and more specifically, our learnings from integrating NIST's AI Risk Management Framework (AI RMF) 1.0 into our platform. We believe that NIST should prioritize work to establish specific definitions and benchmarks for generative AI safety. Over time, NIST can refine its frameworks to make them adaptable based on the size and reach of the implementing organization and its use cases. Our response is focused around five key learnings:

1. **Challenges to AI RMF 1.0 Implementation**. Organizations implementing the AI RMF 1.0 face basic challenges including the need for significant expertise in Responsible AI and require guidance on roles and responsibilities. Providing organization-level requirements and enhancing the AI RMF for specific use cases are suggested solutions.
2. **Guidelines and Standards for AI Safety**. Standardizing disclosures across the AI value chain is crucial. This includes model and dataset cards, conformity assessments, and algorithmic impact assessments (AIAs). These measures will enhance transparency and trust in AI systems.
3. **Disclosure Requirements**. Responsible AI requires disclosures at both organizational and model levels. It is essential that large language and foundation model developers provide governance artifacts and impact evaluations to downstream users for them to successfully identify new risks based on the intended use case.
4. **Transparency and Documentation**. Governance artifacts like model cards, dataset cards, and AIAs are fundamental for transparency. These should be standardized to enhance adoption and interoperability across the AI ecosystem for both traditional and generative AI.
5. **Benchmarks and Evaluations**. Standardizing evaluations –built on industry-driven and consensus-led global standards from international standard-setting organizations like NIST– is key to building a robust evaluation and assurance ecosystem for both traditional machine learning and generative AI.

We believe it is crucial for NIST to prioritize establishing context-focused standards and benchmarks that can help take some of the guesswork out of compliance with AI regulations. We dive deeper into the above learnings by addressing AI safety best practices, guidelines and standards, and benchmarks and evaluations to help inform priorities and next steps for NIST with a particular focus on generative AI.

Best Practices for AI Safety

The E.O. 14110 directs NIST to establish guidelines and best practices to promote consensus-driven industry standards in the development and deployment of safe, secure, and trustworthy AI systems, including developing a companion resource to the AI RMF 1.0 for generative AI, and creating guidance and benchmarks for evaluating and auditing AI capabilities. To help guide our comments in these two areas, the first part of our response focuses on best practices for AI safety based on Credo AI's experience with organizations in terms of "where they stand today," obstacles to implementing NIST's AI RMF 1.0, and underlying changes that AI actors need to make at the organization level to successfully mitigate generative AI risks.

Generative AI Governance Landscape Today

Our experience indicates that most organizations are not ready for generative AI governance given the uncertainty about what changes they need to make to identify and manage risks. The lack of enterprise-ready guardrails to understand and manage risks built into generative AI systems is a key concern for organizations. Many organizations are worried about not having the right tools, expertise, or processes to effectively manage and mitigate the risks associated with using generative AI. While governance maturity is also low for developers and users of traditional ML and AI systems, there are additional barriers to generative AI governance given the pressure to adopt and innovate quickly.

Organizations are concerned about the lack of clarity, guidance and transparency around the use of generative AI internally or within third-party tools. They want to know how to use it, and which generative AI use cases would be considered low risk. Many established organizations, specifically within regulated sectors, want to ensure these guidelines are established prior to allowing generative AI use. Large enterprises will only allow small teams to experiment, or will otherwise limit its use for internal purposes to only enhance existing low-risk functions, such as code checking or creating marketing and design materials.

Integrating third-party foundational models for consumer facing applications requires additional investment and organizational alignment including comprehensive and accurate assessments of AI applications. At Credo AI, we have seen firsthand how transparency and disclosure reporting throughout the AI development life cycle can encourage responsible practices to be cultivated, engineered, and managed.

For meaningful accountability, the AI value chain requires disclosures and transparency throughout the entire "AI stack." Particularly for generative AI, both risk (and conversely, governance) can be introduced at several levels of the "genAI stack." Foundation models — such as GPT-4, Claude, LLaMa, and StableDiffusion — are at the heart of generative AI tools and systems, but they are only one component of the larger generative AI stack.

Generative AI has made the need for AI governance—the coordination of people, processes, and tools to ensure AI risks are effectively mitigated at every stage of the AI lifecycle and at every layer of the AI stack—clearer than ever. Enterprises that want to unlock the value of generative AI need to establish processes to evaluate the risks of the AI applications they're building and buying, adopt controls that mitigate against those risks, and assign accountability for ongoing risk management of AI.

To develop a successful generative AI companion resource to RMF 1.0, as a practical tool and a mechanism for meaningful accountability, it is useful to expand on current obstacles to RMF 1.0 implementation.

Obstacles to Implementation and Solutions

In our experience working with enterprises, some of the obstacles and potential solutions for RMF implementation include:

- Significant expertise in Responsible AI is required to operationalize the NIST AI RMF 1.0. Clearer definition of roles and responsibilities associated with the NIST AI RMF components would be helpful for organizations trying to identify who is responsible for this work—different roles and skills are needed to implement the NIST AI RMF in its current form. Publishing examples, templates, and use-case-specific guidance can also lower barriers to adoption. For example, suggesting individual personas responsible for completing a use case risk assessment or uploading evidence of data governance at various stages of the process would improve both efficiency and accountability.

- Defining risks for an AI use case, ways to measure those risks, measuring them, and taking mitigating actions requires that individuals adopting the RMF have significant knowledge on how to identify, measure, and mitigate AI risks. This can be simplified with tailored versions of the NIST AI RMF that suggest risks and measurement methodologies based on specific use cases (i.e. a credit risk prediction use case is probably going to have similar risks across different organizations).
- More broadly, organizations are not making the needed structural changes required to implement robust responsible AI programs, and fully adopt NIST's RMF 1.0. We dive deeper into this obstacle and propose three changes that AI actors need to make at the organization level in order to effectively adopt RFI 1.0 and consequentially a generative AI companion resource.

Organizational needs for generative AI risk management

To confront the challenges exacerbated by generative AI, including confabulations, harmful content, algorithmic bias, misinformation, intellectual property, privacy, cybersecurity, and other unknown risks, it is essential for organizations to:

- **Assign individuals** within the organization to take accountability for managing these risks and most importantly to equip them with the necessary resources and authority;
- **Establish clear policies**, processes, and a dedicated program to define acceptable uses of generative AI while identifying potential risks and taking proactive measures to address them; and,
- **Provide comprehensive training** to empower the organization's workforce to understand the capabilities and limitations of generative AI. This fosters responsible usage and minimizes risks, ensuring an organization utilizes generative AI effectively while maintaining a secure and well-managed environment.

From our experience working with enterprises to operationalize AI governance, the types of professions, skills and expertise organizations need to effectively govern generative AI include:

- Individuals who deeply understand the business and the intended use (e.g. Product Managers);
- Individuals who can comment on the technical details of the system (e.g. Engineers who developed the models that make up the system);
- Individuals who understand the legal dimension with experience in compliance, civil liability, and contracts, (e.g. Product Legal Counsel);
- Individuals with cybersecurity expertise (e.g. Trust & Safety and Engineering Leads); and,
- Individuals who bring these often siloed teams together (e.g. Technical Program Managers).

Organizations are interested in standards and assessments that help them control risks such as toxicity, hallucination and accuracy of their models. However, some organizations approach generative AI governance from an organization-level or design-guidelines perspective only. While organizational level changes as outlined above are imperative, it should not preclude adopting a systems-level governance approach.

Guidelines and Standards for AI Safety

To achieve transparency in AI, it is necessary to inject disclosures throughout the entire AI value chain. Disclosures can take the form of several different types of governance artifacts, including model and dataset cards, conformity assessments, and algorithmic impact assessments (AIAs). It is important to consider requirements for private companies to provide governance artifacts that include both risk- and impact-informed evaluations of their AI system, and transparency into the governance actions organizations are taking pre- and post-deployment.

Disclosure Requirements

A baseline of Responsible AI practical requirements is required to safely develop and adopt generative AI. Large language and foundation model providers should be required to:

- **Provide organizational-level process (and use) disclosures**. Organizations should be required to disclose (to the general public) if and how they are using LLMs/FMs. Beyond transparency disclosures at the model level, LLM/FM providers must institute organization-level policies for responsible use and development, including investment in AI safety and governance, and processes to ensure full lifecycle governance.
- **Provide model-level transparency disclosures**. Such disclosures should be made available to downstream application developers (not the general public) to ensure they are fit for purpose. Disclosures can take the form of several different types of "governance artifacts" including comprehensive AI use case risk reporting, model cards, data set cards, and AIAs described in detail below. Information about red-teaming methodologies and outcomes can increase the robustness of these disclosures.

AIAs are indispensable transparency tools for entities throughout the entire AI value chain, from foundational model providers to the developers who build upon them, to the end users using AI in their daily lives. The OMB Draft Guidelines on "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" already includes AI Impact Assessments as minimum practices for safety-impacting and rights-impacting AI procured by federal agencies; this should become an industry standard.

Transparency and Documentation

Fundamentally, governance artifacts are the building blocks of trust in the AI Ecosystem as a whole. **By standardizing disclosures, the entire value chain can be accelerated, speeding the application of AI technologies in various industries**.

- **Model Cards** should document basic technical information about an ML model that is relevant regardless of where and how the model is going to be deployed. Model Cards can be created by model builders before a final end application for the model has been identified. Model Cards are designed to give stakeholders visibility into how a model works and what it does, along with potential technical limitations and risks.

- **Dataset Cards** are designed to provide some basic information about what a dataset is and contains, its source, and its potential limitations for use in training (or testing) an ML system. Dataset Cards provide critical supplementary transparency about the factors that influenced a model's development and facilitate third-party auditing and reproducibility.

- **Conformity assessments** enable buyers, sellers, consumers, and regulators to have confidence that products sourced meet specified standards. Conformity assessments ensure that products, services, systems, persons, or entities meet certain required characteristics, and that these characteristics are consistent from product to product, service to service, system to system, etc.

- **Algorithmic Impact Assessments** include metrics related to the performance, robustness, and fairness of an AI system, explainability documentation, and an assessment of risks from intended and unintended use. AIAs are already a widely known and accepted form of assessing potential risks and impacts in other domains (used for environmental, privacy, cybersecurity, and human rights assessments), and have been shown to support a vibrant innovative ecosystem while supporting safety.

Algorithmic Impact Assessments (AIA)

Some elements of AIAs should be standardized but not all (e.g. some AIAs should be more focused on privacy than safety depending on the organization). Baseline requirements for AIAs across all agencies should include:

- Use Case Description
  - Application Context
  - Task & Output
  - Data & Input
  - ML Models (with model cards)

- Responsible AI Evaluation
  - Abuse & Misuse: How well is the system protected from the potential to be used maliciously or irresponsibly (intentionally or unintentionally)?
  - Performance Evaluation: How well does the system perform, what are the accuracy and performance limitations?
  - Fairness Evaluation: Is the system causing disparate impact or harm to different groups that interact with it?
  - Security Evaluation: Is the system secure against adversarial attacks or other bad actors' manipulations?
  - Privacy Evaluation: Does the system preserve individual privacy?
  - Explainability and Transparency Evaluation: Is the system understandable and transparent to all relevant stakeholders?
  - Environmental & Societal Evaluation: What broader changes might the AI system induce in society and the environment, such as labor displacement, mental health impacts, or the strain on natural resources and carbon emissions caused by training complex AI models?

- Risk & Harms
  - Assessment: What are the potential risks and harms caused by the system. What is the likelihood and impact of each risk and harm?
  - Mitigation: What steps have been taken to mitigate the identified risks and harms?

AIAs should also consider and include:

- Context and use case specific information tailored to meet the needs of each agency.
- Information that helps determine if a model is fit for purpose.
- Information about what non-AI solutions were explored, the relevant risks and benefits for that solution, and why a particular AI system was ultimately chosen.
- Descriptions of why certain performance and evaluation metrics were selected.
- Reports that are readable by employees and data scientists at each agency.
- Documentation on plans to proactively identify and address new risk scenarios.

Liability is distributed throughout the AI value chain; as such, creating and providing transparency documentation is the responsibility of both providers and downstream developers of AI. Examples (or explicit references to templates) on the elements that should be included in these AI model cards and dataset cards would further enhance the usefulness, viability, and interoperability of these transparency mechanisms.

Many organizations are reluctant to share results about the behavior of their AI systems externally—because they have no idea how their results might compare with those of their competitors, or whether they are “good” or “bad” for external stakeholders. We are strong supporters of reporting requirements, therefore, that promote and incentivize public disclosure of AI system behavior and operation as a key driver of the establishment of standards and benchmarks.

Downstream developers can only perform “last mile” due diligence if they have the necessary documentation to understand how a foundational generative AI model was trained and decide what additional safeguards they need to implement to identify and mitigate use case specific risks. We believe that the practices outlined above will result in positive downstream effects that encourage the adoption of Responsible AI practices in industry and commercial sectors, similar to how government procurement standards have led to the widespread adoption of data encryption and energy efficient building standards.

Benchmarks and Evaluations

The availability of benchmarks that can reliably indicate the impact of generative AI and LLM agents on a very large scale is limited. To generalize standards and methods of evaluating AI over time that are interoperable across sectors, and across use cases, is challenging. Baseline model benchmarking and testing and structured mechanisms for gathering human feedback are ideal, but still face significant limitations.

Rigorous evaluation of capabilities and limitations should be required for both system developers and users of an AI system. Responsibility is everywhere throughout the AI value chain, although liability can be more narrowly defined. All entities throughout the AI value chain have some capacity for creating a safety review, and the information included in that will depend on their capabilities. Standardizing evaluations is key to building a robust evaluation and assurance ecosystem. Investing in improved evaluations on a host of dimensions (like propensity for misinformation) can help control and align AI systems, providing the infrastructure for a proactive accountability system.

Test Environments - Generative AI Sandboxes

To safely explore generative AI capabilities, the enterprise requires a secure, risk-managed environment (or sandbox) to use in experimentation and proof-of-concept activity, as well as ad-hoc employee use of generative AI tools. The purpose of a Generative AI sandbox is to enable organizations to pilot Generative AI applications and uses of generative models to then implement those applications and models outside of the sandbox safely.

In the context of generative AI, controls are configurable instruments that edit model input or output to reduce risk to acceptable levels, and ensure that the model is functioning according to business requirements – for example, at an acceptable level of efficacy.

Generative models may produce a number of variant responses from a large number of prompts, which means that coverage of every output is impractical, and up front identification of controls is challenging. Some controls need to be discovered through examination of the use and output of generative models. Discovery of controls is achieved through two processes, namely: (i) identification of risk levels associated with different use cases, and (ii) discovery and configuration of controls that reduce risk to an acceptable degree for each use case.

A generative AI sandbox allows risk management stakeholders to identify the relative effectiveness and risk of different approaches. This can be achieved through exploratory analysis or through unsupervised learning techniques such as clustering of inputs. Use cases may be identified through analysis of clustered inputs, and risk and effectiveness levels may be assigned to use cases accordingly.

A sandbox also allows for controls to be implemented, tested, and verified. A control on an LLM might be a filter on inputs (for example, a filter identifying copyrighted material), or a filter on output (for example, a filter identifying bias).

Once a control and associated filters are defined, it is possible to observe the effect of that control on the model's inputs and outputs and therefore evaluate the efficacy of the filter. Subsequently, the control's efficacy in reducing the risk associated with a use case can be measured.

Conclusion

It is crucial for NIST to prioritize establishing context-focused standards and benchmarks—that are globally interoperable—to take out some of the guesswork currently encountered by organizations seeking to adopt responsible AI practices. Without clear standards and benchmarks, organizations are left having to develop and justify *their own* measures for different technical dimensions of their AI systems.

While many emerging regulations set "fairness," "transparency," and other Responsible AI dimensions as key requirements for compliance, there are not yet clear standards or benchmarks for what it means for an AI system to be "fair" or "transparent." A priority for NIST should therefore be to define these terms with language that can be technically implemented by engineering teams, and suggest quantitative benchmarks as initial targets based on broad categories of intended use. Eventually, such broad categories can be refined into profile-based frameworks with use case specific thresholds. While these thresholds would need to be periodically revised and updated, having a baseline could significantly reduce existing uncertainty for enterprises about what AI governance means *in practice*. Standards and benchmarks should also try to account for the challenges of operationalizing such requirements depending on the size and reach of the organization.

Credo AI appreciates both the opportunity to comment on these issues, as well as NIST's highly commendable and continued efforts to advance AI governance and risk management. We look forward to continuing to partner with NIST to help complete obligations under Executive Order 14110 to Support Safe, Secure and Trustworthy Development and Use of Artificial Intelligence.

We welcome any further opportunity to provide resources or information to assist in this important effort. If you have any follow-up questions regarding these comments and recommendations, please contact Credo AI's Director of Global Policy, Evi Fuelle (evi@credo.ai).

Sincerely,

Navrina Singh

Founder and CEO
Credo AI
www.credo.ai