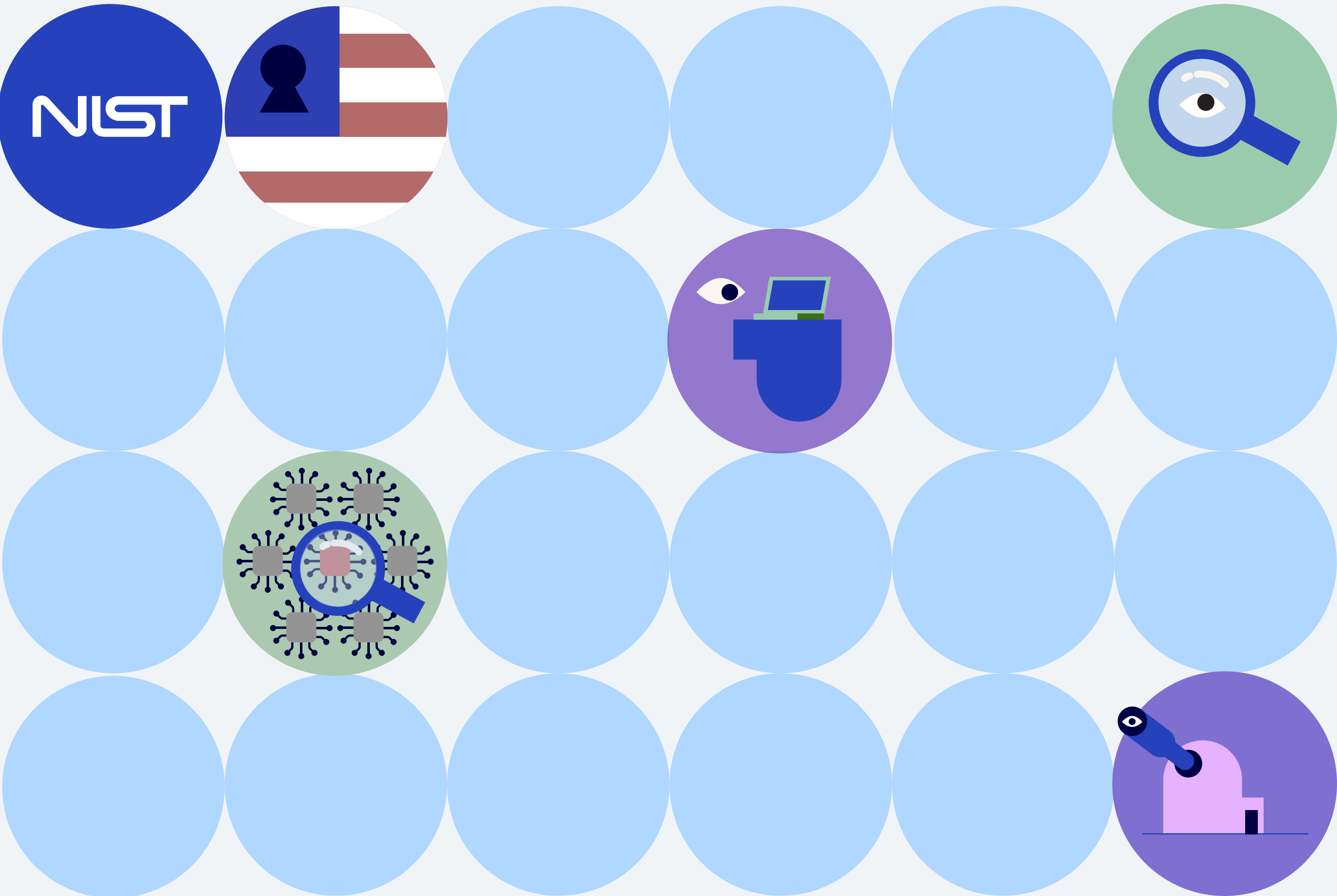
Program Report 2

Open Loop | Meta

Generative AI Risk Management and the NIST Generative AI Profile (NIST AI 600-1)



Contents

Foreword – Meta 03

Erin Egan
Vice President Privacy Public Policy
and Chief Privacy Officer, Meta

Foreword – Accenture 04

Arnab Chakraborty
Chief Responsible AI Officer, Accenture

About Open Loop 05

Executive summary 09

1 Introduction 15

Focus of Phase 2: NIST's Generative AI Profile and related AI guidance 17

Methodology 18

How to read this report 19

Contents

2 **Cross-cutting issues in generative AI risk management** 21

2.1 The AI value chain, actors, roles and responsibilities 22

2.2 Open-source AI guidance 29

2.3 Interoperability of AI Risk Management Frameworks 34

2.4 Evaluating the effectiveness of the NIST AI RMF 1.0 37

3 **Findings & policy recommendations specific to NIST's Generative AI Profile 600-1** 38

3.1 List of generative AI risks 39

3.2 Usability and practical implantation of the guidance 43

Conclusion 45

Annex 1 - Methodological Note: Open Loop US Program Phase 2 - NIST's AI RMF 1.0 Adoption and draft Generative AI Profile Feedback 48

Foreword

I am excited to be sharing with you the second report from our Open Loop US program on generative AI risk management. In setting out to deliver this program we had a central goal to work with people and companies to inform the principles which will guide the safe and responsible development of generative AI now and into the future. This is crucial work which requires investment, cutting-edge research, and deep collaboration.

When we kicked-off the program last November, President Biden had just signed the Executive Order on "Safe, Secure and Trustworthy Development and Use of Artificial Intelligence", which directed NIST to produce companion guidance to the AI RMF 1.0. This guidance was to drive consensus in industry standards, to be completed in broad consultation with stakeholders, and delivered within an ambitious timeframe. With this report we are pleased to recognize the work of NIST in quickly mobilizing to meet this challenge. In July 2024 NIST published its generative AI-focused guidance, or "Profile" — NIST AI 600-1 — alongside other important complementary documents on safe AI software development and international cooperation. These documents form a stable yet flexible architecture for generative AI risk management and international cooperation towards AI safety standards.

Due to the rapid development of this guidance, in this second phase of the Open Loop program we have had the opportunity to focus on gathering direct responses from companies to the Generative AI Profile as well as the AI RMF 1.0, and we thank the companies participating on this phase of the program for pivoting nimbly to support this undertaking. We are indebted also to the continued contributions of our group of experts who have generously assisted us in our efforts to ensure that the program keeps pace with AI policy and technology developments.

Over years of running our Open Loop programs around the world we have seen first hand that gathering structured, in-depth feedback from companies as they respond to legal or regulatory guidance yields a wealth of tangible suggestions for improvements to the guidance in question — enhancing its clarity, actionability and efficacy. We look forward to continuing this journey with the community and further supporting the advancement of responsible generative AI policy and practice as this technology reaches exciting new heights.

Erin Egan
Vice President Privacy Public Policy
and Chief Privacy Officer, Meta

Foreword

The rapid evolution of artificial intelligence (AI) technologies, particularly generative AI, has brought about both significant opportunities and some unique challenges. It is changing the way we work and live, but it also introduces new risks and regulatory dynamics. Accenture research shows that 96% of companies support some kind of government rules about AI. Currently 2% of companies say they have *fully* operationalized responsible AI in their companies, however 31% expect to do so in the next 18 months.

Accenture has helped dozens of clients begin the process of integrating NIST's AI RMF and now, the Generative AI Profile, into their own practices. We have also embedded this foundational guidance into our own business through our Responsible AI Compliance Program, which helps our teams use AI effectively. We brought those experiences to our research with program participants and development of recommendations for this report.

When we began working with Meta to deliver the Open Loop US program in November 2023, we were excited for the opportunity to use our broad experience with AI actors across the value chain to better understand where companies are struggling to implement NIST's guidance.

Through the program and together with companies, academics and practitioners we have identified numerous opportunities for NIST to enhance their guidance. Areas which call for particular attention include AI value chain transparency and risk management for open source models, as well as pragmatic aspects of policy innovation that will make it easier, faster and more efficient for organizations to implement.

These findings align with Accenture's experience on the ground helping clients rapidly mobilize generative AI technology while attempting to avoid the complex issues that often accompany emerging technologies. In our work with clients, we see that successful management of generative AI-related risks is a shared effort between multiple parties. Initiatives such as Meta's Open Loop programs are critical for fruitful cross collaboration on generative AI risk management.

We hope the insights provided here help NIST, and all industry actors take the next right step in their journeys. We are not tackling these challenges alone. I look forward to future collaboration efforts as we strive to realize AI's potential to transform how we work and live, creating better societies for all.

Arnab Chakraborty
Chief Responsible AI Officer, Accenture

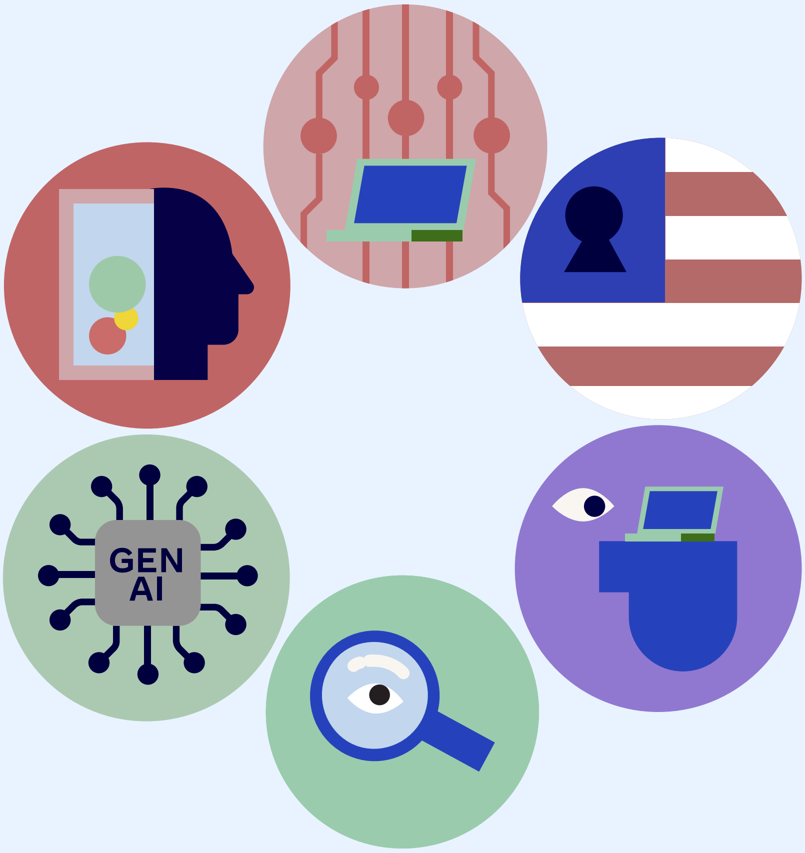
About Open Loop

Meta's Open Loop is a global program that connects policymakers and technology companies to help develop effective and evidence-based policies for AI and other emerging technologies through gathering feedback on new or existing policies, regulations, laws, or voluntary frameworks.

The aim is to improve the quality of guidance and regulation on emerging technologies, ensuring that they are understandable, feasible in practice, and likely to affect the intended outcome of the policy.

This report presents the findings and recommendations of the second phase of the Open Loop US program on generative AI risk management. We launched this phase in July of 2024, and again in partnership with Accenture, as with the first phase.

How to cite this report?
Galindo, Laura; Naidoo, Taja; Nugteren, Maartje; and Shah, Ali. "Open Loop US Program Report 2: Generative AI Risk Management and NIST AI 600-1", (November 2024), at https://openloop.org/reports/2024/10/Report-2-NIST-Generative-AI-Profile.pdf



Acknowledgements

The Open Loop US program was created by Meta, and designed and implemented in collaboration with Accenture. Our thanks to Ali Shah, Rachel Pelham, Patrick Connolly, Haley Rosowsky, Shekhar Tewari, and Jakub Wiatrak and Mariah Barber for their contributions to the second phase of the program.

We would like to thank the following organizations for their participation in the program — without their commitment and active involvement, this work would not have been possible:

- Active Fence
- Applause App Quality Inc
- AI Collaborator Inc
- Avanade, incl.
- Credo AI
- Direct Relief
- Holistic AI
- LegalMente AI
- Local Spark AI
- Numbers Protocol
- RWS Group
- Velocity Global
- VU Studio
- Zaviant

One company who completed the survey preferred to participate anonymously.

Acknowledgements

We are indebted to our group of experts from across the AI ecosystem who shared their deep expertise, contributed to the development of the program's research strategy, and supported knowledge building amongst the company cohort:

- Dr. Anthony Barrett | Center for Long-Term Cybersecurity, UC Berkeley
- Benjamin Larsen | World Economic Forum
- Carlos Ignacio Gutierrez |The Future of Life Institute
- Carlos Gutierrez | LGBT Tech
- Cobun Zweifel-Keegan | International Association of Privacy Professionals
- Courtney Lang | Information Technology Industry Council
- Cristina Pombo | Inter-American Development Bank
- Daniel Castro | Information Technology & Innovation Foundation
- Elena Estavillo | Centro-i for the Society of the Future
- Eli Sherman | Credo AI
- Dr. Ellie Graeden | Georgetown University Massive Data Institute
- Eugenio Zuccarelli
- Evi Fuelle | Credo AI
- Dr. Heather Frase | verAItech; Virginia Tech, National Security Institute
- Jennifer Jin | Centre for Information Policy Leadership
- JudeAnne Heath | HTTP (Hispanic Technology & Telecommunications Partnerships)
- Jutta Williams | Humane Intelligence
- Kristian Stout | International Center for Law & Economics
- Laila Abdelaziz | Centre for Information Policy Leadership
- Lucia Gamboa | Credo AI
- Luciano Floridi | Digital Ethics Center, Yale University
- Markus Heyder | Centre for Information Policy Leadership
- Marlena Wisniak | European Center for Not-for-Profit Law
- Matt Mittelsteadt | Mercatus Center
- Matthew Reisman | Centre for Information Policy Leadership
- Miranda Bogen | Center for Democracy & Technology
- Neil Chilson | Abundance Institute
- Nile Johnson | Institute for Security and Technology
- Philip Dawson | Armilla AI
- Dr. Rumman Chowdhury | Humane Intelligence
- Stephanie Ifayemi | Partnership on AI
- Tom Romanoff | Bipartisan Policy Center

We would also like to express our gratitude to the observer institutions that participated as part of our expert group: the Info-Communications Media Development Authority, Singapore (IMDA); the Department of International Law, Organization of American States (OAS); the National Data Protection Authority, Brazil (ANPD); the United Nations Institute for Disarmament Research (UNIDIR); and the Organization for Economic Cooperation and Development (OECD).

Finally, thank you also to our design partners at Craig Walker Design and Research who helped us transform our data into a well-designed report.

Executive Summary

This is the second and final report from the Open Loop US program on Generative AI Risk Management which looked in detail at NIST's AI RMF 1.0 and subsequent supporting documents.

> The aims of the program were to inform participating companies about the AI RMF 1.0, to learn about their current generative AI risk management practices, and, in the second phase, to gather feedback from these companies specifically on NISTs draft "Generative Artificial Intelligence Profile" (NIST AI 600-1).

Here we present a high-level summary of the findings and recommendations from the second phase of the program. We found that:

- ✓ Overall, the companies welcomed NIST AI 600-1 (The Generative AI profile), with the majority of organizations agreeing that it "complements and enhances" the AI RMF 1.0 through the provision of specific guidance on addressing generative AI risks;
- ✓ However companies still need more guidance on developing, optimizing and deploying generative AI in order to best balance safety, privacy and innovation.

Specific challenges cited were as follows:

→ Most companies were still “unclear” or “very unclear” about their position and role in the AI value chain, but felt that this was important to establish, as it influences their risk management responsibilities and prioritization;

→ While the majority of organizations and experts find the 12 generative AI risks in the Generative AI Profile comprehensive, there is a strong demand for more nuanced categorization, context and sector-specific guidance, and consideration of risk interactions;

→ Over two thirds of respondents find guidance on open-source approaches to generative AI risks very or extremely useful, indicating an opportunity for NIST to provide more detail on open-source spectrum release strategies;

→ 90% of organizations expect implementing the Generative AI Profile to create a “high” or “very high” workload with limited staffing (73%) and complexity of guidance (55%) as top barriers for implementing the Generative AI Profile.

→ Virtually all companies reported difficulty navigating multiple AI governance frameworks. Despite current efforts by a number of organizations (including NIST) to create alignment, greater framework interoperability regarding generative AI risk management is needed. Further, experts highlighted the potential of the AI Safety Institutes to aid global harmonization of standards and practices, in generative AI development and deployment.

From these findings, we have formulated the following recommendations to NIST:

1 Expand and codify the AI value chain: actors, roles and responsibilities.

NIST should conduct research pursuant to updating the taxonomy of AI Actors, perhaps in cooperation with the OECD, so that it better reflects the complex dynamics and relationships within the developing generative AI ecosystem. Additionally, acknowledging the responsibilities of all actors in the AI value chain and seeking to ensure clarity and consistency across guidance frameworks, NIST should reframe the subsequent draft guidance in NIST AI 800-1 to clarify that suggested actions apply to both model providers/developers and deployers, and correspond to their role or roles in the AI value chain.

To support the transfer of useful information and enhanced transparency among actors NIST could consider facilitating a collaborative process with the AI community to develop a set of templates for information sharing up and down the AI value chain which balance transparency, comprehensiveness and confidentiality and ensure compliance with existing laws, perhaps leveraging work already done by Partnership on AI on Machine Learning documentation.

Finally, NIST could explore the creation of a voluntary pilot program, “sandbox” or other research-focused program where different actors in the AI value chain can test and refine these information-sharing practices with real use cases in a secure environment. Some of these practices could include co-developing templates for Model Cards or “Fact Sheets”, or developing novel types of documentation.

2 Provide detailed, consistent guidance specific to the variety of open-source approaches and contexts.

NIST should produce a set of suggested risk management actions tailored to open-source AI approaches and release strategies, guidelines on secure data handling and model fine-tuning in open-source spectrum contexts, as well as strategies for adapting to rapid innovation in the AI ecosystem. It should also seek to align guidance on risk evaluation and measurement for Foundation Models — specifically around the utility of marginal risk — with The National Telecommunications and Information Administration's (NTIA) guidelines for "Dual-Use Foundation Models with Widely Available Model Weights" issued earlier this year to avoid confusion among companies.

Lastly, NIST should develop a "quick start" guide based on the AI RMF and Generative AI Profile, but aimed at smaller companies (SMEs and startups) — this should provide streamlined guidance for these companies to embark upon open-source generative AI responsibly, including a step-by-step overview of risk assessment and mitigation which links out to more detailed resources including the full NIST AI 600-1 and AI RMF 1.0 documents.

3 Further develop an interactive tool or "alignment map" which would map principles and practices across the most used generative AI risk management frameworks globally.

NIST should continue its efforts in developing comprehensive "crosswalks" between major AI governance frameworks, coupled with interactive tools to help companies navigate requirements based on their specific context and sector. This would help companies to understand where requirements or recommendations overlap and ensure efficiency of integration within a single risk management system. Frameworks which should be prioritized for these mapping efforts include the EU AI Act and ISO/IEC 23894:2023.

NIST should specifically consider aligning its risk taxonomy more closely with emerging international frameworks which focus on transparency and reporting, such as the OECD's work on AI incidents and hazards.

4 Consider establishing a dedicated workstream to co-design and explore methodologies for assessing the RMF's effectiveness with the broader AI ecosystem.

This effort could build upon exercises like the Open Loop program, which provide empirical data on the framework's real-world application. Such a workstream could develop metrics to measure the framework's impact more holistically, establish mechanisms for regular user feedback, conduct longitudinal studies, collaborate on independent evaluations, and create a public repository of case studies and best practices. NIST can ensure through this process that the framework continues to evolve in response to the rapidly changing AI landscape and the needs of its users. This approach aligns with NIST's commitment to maintaining the AI RMF as a "living document" and would provide valuable empirical evidence to inform future updates and refinements. In doing so, NIST can ensure that it remains a relevant and impactful tool for managing AI risks across diverse contexts and applications.

5 Adapt the list of risks in the Generative AI Profile into a dynamic, interrelational risk matrix with sub-profiles for specific sectors and aligned guidance on how to measure and evaluate risks.

NIST should iterate on this list and create a relational framework that acknowledges the interconnected nature of risks, showing how they interact and might potentially compound each other, and providing examples. Due to the diverse nature of AI developers and deployers and the salience of context and use case, NIST could create sector-specific “sub-profiles” and real examples to help organizations translate generic risk descriptions into actionable insights for their particular domains, for example in the manufacturing or healthcare sectors.

NIST should also clarify the distinction between risks, hazards, and harms. Clear definitions and examples should be provided to help organizations differentiate between these concepts.

NIST could further integrate guidance on anticipating and preparing for emerging and long-term risks.

⑥ Consider the usability (or UX) elements of the Generative AI Profile and companion guidances, focusing on enabling small companies to efficiently find and understand their responsibilities.

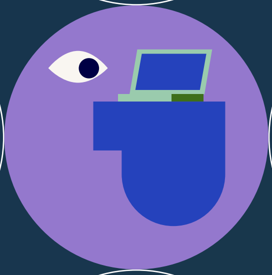
NIST could develop — over time and in collaboration with industry working groups — a tiered guidance system that caters to organizations of different sizes and generative AI maturity levels. This could include a reader guide — akin to the AI RMF Playbook — for smaller organizations or those new to AI risk management which organizes actions based on actors and position in the value chain, rather than being organized by framework pillars.

In addition, NIST could create online tools and video explainers to guide users through the risk assessment process, helping them identify which parts of the profile are most relevant to their specific context, or even a "self-assessment" simulator for organizations to measure their compliance/adoption levels.

CHAPTER ONE

Introduction

1



This is the second of two reports which share the findings and recommendations which have emerged from the US Open Loop program on Generative AI Risk Management. This program commenced in January 2024 and focused on analyzing the AI RMF 1.0 (Risk Management Framework) produced by the National Institute of Risk Management (NIST) in the United States of America and its application to generative AI with a cohort of US-based companies.

The program overall has two main objectives: to help the participating companies understand NIST's AI RMF and how it can structure and support their risk management efforts, and, to produce valuable feedback on the framework which can help NIST better tailor it to generative AI applications. While this feedback was gathered with NISTs future efforts in mind, we hope that the findings are useful for a broad spectrum of policymakers and indeed any stakeholder who is interested in generative AI risk management.

> We published the first program report on 6 June 2024. That report focused on red-teaming and synthetic content risk mitigation practices, but also established many of the themes which we have further developed in this report. Therefore, you may find it useful to return to that first report. In addition, we recommend reading the guidance issued by NIST in April and July 2024. (See Table 1 below)

Focus of Phase 2: NIST Generative AI Profile and related AI guidance

On 29 April 2024 NIST — in satisfaction of their obligations under the White House Executive Order on Safe and Trustworthy AI[1] — issued four draft reports[2] which complement the original AI RMF 1.0 including a new "Profile" for generative AI which provides guidance specific to mitigating risks arising from the development or use of this technology. They opened a comment period on this document between 29 April and 2 June 2024.

Phase 2's focus: NIST Generative AI Profile and related AI guidance

On 29 April 2024 NIST — in satisfaction of their obligations under the White House Executive Order on Safe and Trustworthy AI — issued four draft reports which complement the original AI RMF 1.0 including a new "Profile" for generative AI which provides guidance specific to mitigating risks arising from the development or use of this technology. They opened a comment period on this document between 29 April and 2 June 2024.

On July 26th 2024 NIST published the finalized version of this Profile. NIST AI 600-1 ("AI RMF Generative Artificial Intelligence Profile", hereafter "Generative AI Profile" or "GenAI Profile"). It centers on a list of 12 risks and mitigating actions that AI actors can take, and aims to assist organizations in managing AI risks in a manner aligned with their legal and regulatory requirements, goals, and risk management priorities.[3]

As the second phase of our program was executed concurrently with NIST's work on the Generative AI Profile, we pivoted quickly from a broader focus on the AI RMF 1.0 and its application to generative AI and focused on gathering feedback on the Generative AI Profile directly. It is important to clarify however that the input we received from companies in the period from early June to early July 2024 — which form the basis of the recommendations in this report — was in response to the DRAFT version of the Generative AI Profile. They have subsequently published the final version of the GenAI Profile, which incorporates feedback received from the AI development community and other key stakeholders.

Table No. 1 — NIST AI draft and final guidance documents

Document	Draft Published	Final Published
Directly addressed in this report through research questions:		
NIST AI 600-1: AI RMF Generative AI Profile	29 April 2024	26 July 2024
Not directly addressed, but relevant:		
NIST Special Publication (SP) 800-218A: Secure Software Development Practices for Generative AI and Dual-Use Foundation Models	29 April 2024	26 July 2024
NIST AI 100-5: A Plan for Global Engagement on AI Standards	29 April 2024	26 July 2024
NIST AI 800-1: Managing Misuse Risk for Dual-Use Foundation Models	26 July 2024	TBC

In the recommendations we have provided in chapters 2 and 3 we acknowledge that NIST has made great strides in addressing some of the original gaps in the draft guidance with their final published version of the Generative AI Profile, however there are still opportunities to be realized which they should consider addressing in future updates to the guidance.

Methodology

Our research involved a comprehensive survey of 15 companies across the AI value chain, in-depth interviews with 14 organizations, and three expert focus groups (see Annex 1 for further details).

Cohort's profile and RAI's practices

Our Phase 2 cohort represented a diverse cross-section of the AI ecosystem, offering insights from various perspectives within the AI value chain. The majority (67%) of participating organizations identified as downstream deployers of GenAI models, while 47% were developing GenAI-powered tools, and 33% were model acquirers. This distribution reflects the current landscape where many organizations are actively integrating and adapting pre-existing AI models rather than developing them from scratch.

The cohort demonstrated a clear trend towards implementing Responsible AI (RAI) governance practices, with 40% actively rolling out RAI-specific practices and another 27% in the formulation stage. This high level of engagement with RAI frameworks and practices aligns with the increasing recognition of the importance of AI governance among organizational leaders. Notably, the cohort included a balanced mix of company sizes and maturity levels, with 46% large enterprises (>250 employees) and 39% small to micro enterprises (<50 employees). The majority (77%) had been established for over 5 years, suggesting a blend of experience and fresh perspectives in our sample. This diverse profile enabled us to gather insights on NIST AI RMF implementation challenges across different organizational contexts and AI value chain positions.

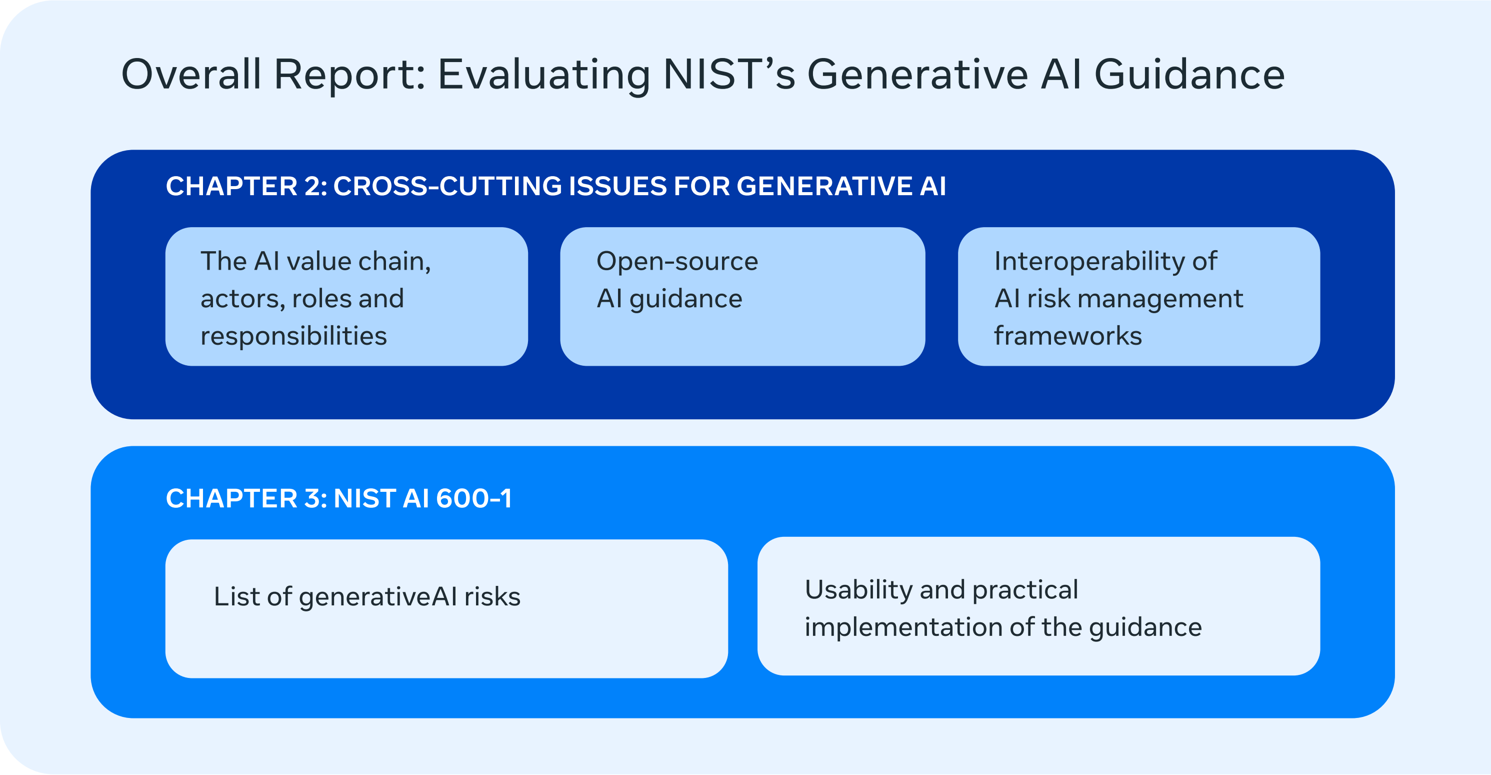
See Annex 1 below (Methodological Note) for more details.

This report synthesizes our findings, highlighting common themes and important opportunities for improvement. It aims to provide policymakers, regulators, standard setting organizations and AI practitioners with an up-to-date understanding of the current state of AI risk management practices and the effectiveness of NIST's guidance in the form of the broad risk management approach described within the AI RMF 1.0 and the specific (draft) provisions within the Generative AI Profile.

How to read this report

This report is divided into two sections:

Overview of Findings & Recommendations – Phase 2



Overall Report: Evaluating NIST's Generative AI Guidance

CHAPTER 2: CROSS-CUTTING ISSUES FOR GENERATIVE AI

The AI value chain, actors, roles and responsibilities

Open-source AI guidance

Interoperability of AI risk management frameworks

CHAPTER 3: NIST AI 600-1

List of generativeAI risks

Usability and practical implementation of the guidance

Chapter 2 focuses on three key areas where NIST could enhance its overall generative AI guidance:

- **Value chain responsibilities:** Clarifying roles and responsibilities throughout the AI lifecycle.
- **Open-source AI ecosystem:** Addressing the implications of open-source AI for risk management.
- **Interoperability and standardization:** Promoting compatibility and alignment between different AI frameworks and standards.
- **Evaluating the effectiveness of NIST's AI RMF 1.0:** Exploring methodologies for assessing the RMF's effectiveness with the broader AI ecosystem.

These issues are cross-cutting and will likely require collaboration across agencies and organizations to address, but NIST has a substantive and influential role to play on each. Due to the complex and interconnected nature of the issues they will also need to be covered across multiple guidance documents outside of the Generative AI Profile, primarily those listed in Table No. 1 above.

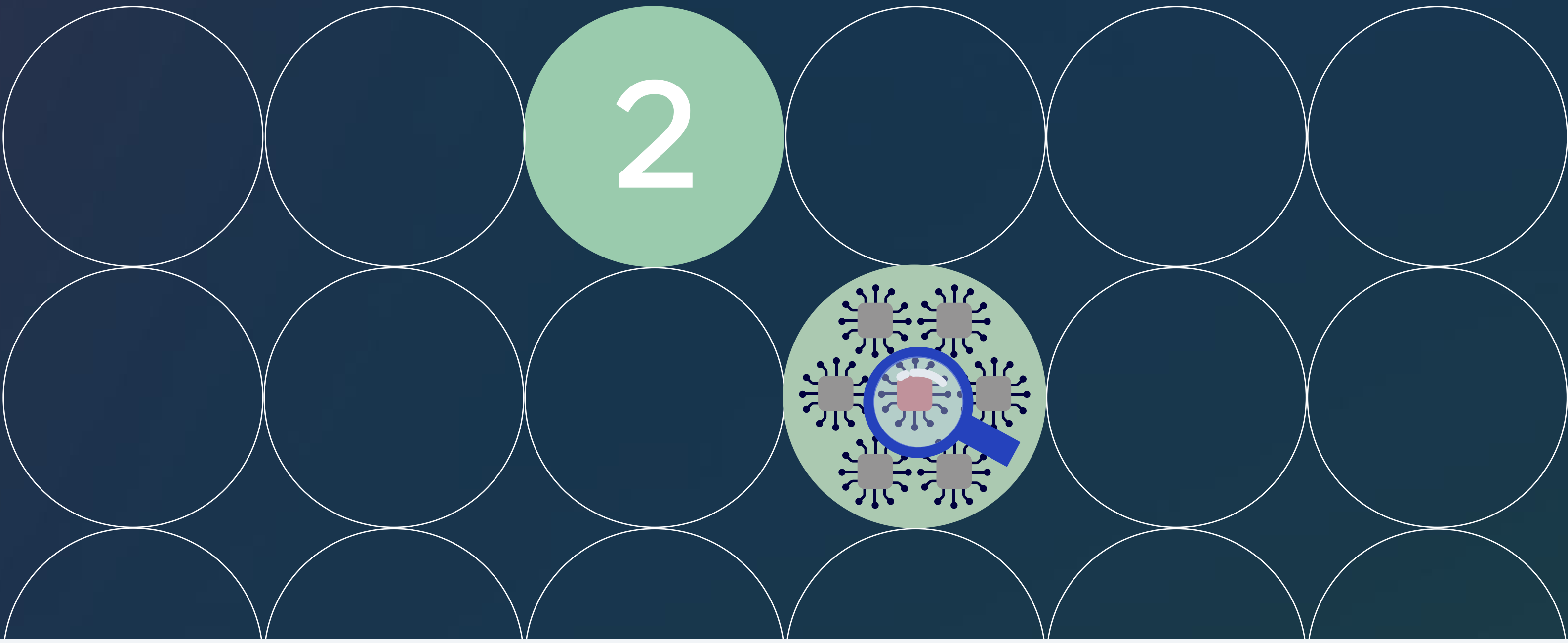
Chapter 3 dives deeper into the NIST 600-1 Generative AI Profile specifically, offering targeted feedback on:

- **List of AI risks:** Evaluating the comprehensiveness and relevance of identified risks.
- **Suggested actions:** Assessing the practicality and effectiveness of proposed mitigation strategies.

As the AI landscape continues to evolve at a rapid pace, we believe this report will serve as a valuable resource for NIST and the broader AI community in refining and enhancing generative AI governance and risk management practices. We invite readers to engage critically with our findings and recommendations, and to consider how they might be applied in their own organizational contexts.

CHAPTER TWO

Cross-cutting issues in generative AI risk management



This chapter lays out the challenges foreseen by companies as they assessed the implications of the draft NIST Generative AI Profile 600-1 (April 2024 version) and other relevant guidance, specifically “NIST AI 100-5: A Plan for Global Engagement on AI Standards” and “NIST Special Publication (SP) 800-218A: Secure Software Development Practices for Generative AI and Dual-Use Foundation Models”.

While these documents have laid an important foundation for generative AI risk management, gaps remain in some areas — AI value chain, open-source guidance and interoperability of frameworks. The issues described are broad, complex, and require coordination between agencies and nations to solve, as opposed to being within the sole purview of NIST to advise on. However, NIST can continue to play a highly influential role nationally and internationally in these areas as they build upon their impressive work to date and continue their cooperative efforts with the OECD, EU-US Trade and Technology Council and the AI Safety Institutes, among others.

2.1 Challenges to conducting stakeholder engagement

Our company cohort uniformly asserted that an organization’s place and role within the AI value chain is integral to determining the risk management responsibilities of that organization and executing them efficiently. This presents two challenges — firstly, determining the position of any actor in an increasingly complex ecosystem of involved parties; secondly, focusing on the responsibilities that fall within their remit and ensuring that each actor has the information needed to fulfill their role.

We outline our findings and recommendations on both of these areas below.

FINDING

2.1.1 **Lack of clarity on actors and roles in the AI value chain**

DETAILS

As echoed by both our participating companies and group of experts, the generative AI landscape has evolved into a complex, non-linear ecosystem with interconnected actors performing various roles. This complexity necessitates a nuanced approach to risk management and governance.

Participating companies were distributed along the AI value chain and, in some cases, occupied more than two places within that value chain. As mentioned in the chapter one, the cohort was dividing along the value chain and many occupied more than one role: 13% reported developing generative AI models themselves or “providing” models to other organizations; 33% were acquiring the models from providers, with 67% reporting that they were also engaged in downstream deployment and fine-tuning of models for specific applications. Respondents were also involved in developing software using AI-generated tools and components.

Despite the significant expertise and experience of this cohort many reported that they were still unclear as to which actions suggested in the Generative AI Profile applied to them, and if these measures could — or should — be taken by more than one actor in the AI value chain. When asked how "clear" they thought the draft guidance was in indicating which actions should be undertaken by which AI actors, just under half of the group (46%) responded that they were either "very unclear", "unclear" or "neutral" to this question, underlining the ongoing uncertainty in this area.

On the question of determining the role of an organization within the AI value chain, while NIST provides an adapted list of AI actors in its final GenAI Profile,[4] it may be beneficial to update this for generative AI and expand it to reflect current practices in the field. For example, at this time it could be helpful to define key AI actors such as: foundation model providers, downstream developers, model deployers and end-users. A more detailed breakdown of the AI value chain, as offered by some research institutions, could provide insight into the number of actors now routinely involved in the development, adaptation, optimization, and deployment of generative AI models. Recent cooperative initiatives, such as Partnership on AI's workshop on mapping the AI value chain[5] also demonstrate encouraging progress in this direction.

Table No. 2 — Suggested AI Actors for inclusion by NIST in the Generative AI Profile

AI Actor (for example)	High-level description
Foundation Model Providers	Organizations that train foundational models and may make them available.
Model Hosting Services	Those platforms that make available foundation model and adapted models to downstream developers
Downstream Developers	Those who use or build upon a foundation model to create their own use cases, often fine-tuning the models with their own data.
Model Deployers	Those that integrate AI models into products or services
End-Users	Those who interact with the fine-tuned models or AI-powered applications.

Note: This non-comprehensive taxonomy aims to capture the nuanced roles that have emerged in the generative AI ecosystem. However, it's important to note that these categories are not mutually exclusive, and many organizations may occupy multiple roles simultaneously.

DETAILS

Since the data provided as well as the goals and characteristics of the AI model or system and the context of use may vary greatly between organizations, it would be valuable to establish clarity around the responsibilities of each of these AI actors within the development and deployment chain. NIST has also this summer released draft guidance on “Managing Misuse Risk for Dual-Use 3 Foundation Models” (NIST AI 800-1). Our findings would suggest that this draft guidance will further confound matters with regard to the roles and responsibilities of organizations within the AI value chain, being focused — as it is — solely on foundation model providers. This focus appears to place the document at odds with the AI RMF 1.0[6], the Generative AI Profile[7], and the White House Voluntary Commitments, which all acknowledge the need for a holistic view of the distribution of responsibilities throughout the AI value chain. Rather than being inconsistent with previous guidance and deepening confusion in this area, NIST’s 800-1 document could make a useful contribution to clarifying responsibilities along the value chain by stating that this guidance can apply to both model developers and deployers depending upon their role, as well as by clarifying that the Marginal Risk Framework used in the recent NTIA report “Dual Use Foundation Models with Widely Available Model Weights” is the appropriate framework for analyzing and addressing misuse risk.

Further, acknowledging our finding that many actors occupy more than one role in the value chain, it would also be helpful to clarify how these actors should consider their responsibilities as they move between these roles through the different stages of the AI development and deployment lifecycle, and which of the responsibilities might — or should — also apply to the actors upstream or downstream from them.

More work is needed on the "definitions" of these actors and to reach a consensus view on which additional definitions are meaningful from a risk management perspective. There remains much to be explored, particularly in considering the entire spectrum of model uses, from deployment by model developers to distribution on platforms or hosting intermediaries to specific business uses. Ensuring that the list of actors and definitions remain under review will be important, as this is an evolving area that will require ongoing community engagement and advancement as the ecosystem continues to develop.

FINDING

2.1.2 **Opportunities for enhancing collaboration and transparency across the AI value chain**

DETAILS

Building on the need to clarify roles within the AI value chain, program participants highlighted the importance of defining what information should be shared between different actors, and what responsibilities would sit with each actor. As mentioned at the start of this chapter, the interconnected nature of the AI value chain necessitates a collaborative approach to risk management. Each actor’s actions can significantly impact others down the line. For instance, the safety measures implemented by foundation model providers can be undermined if model adapters or downstream developers are not aware of — or do not maintain — these safeguards.

Companies in the program agreed with this, and pointed to the challenges of executing their responsibilities within the AI value chain in the face of a perceived deficit of information about the generative AI models or systems they are acquiring or optimizing, and, in the case of model providers, a lack of clarity around how end-users may be employing the models. Participating companies talked of the lack of approaved templates which could allow companies to exchange information — particularly regarding security or privacy incidents — more confidently.[8]

While several actions (e.g., MG-3.1-005, MG-3.2-003) in the final Generative AI Profile emphasize the need for transparency artifacts and documentation which apply to both providers and deployers, they do not clearly distinguish which party is responsible for sharing different types of information. Indeed, under GV-2.1-001 the implication is that responsibility for determining "roles, policies, and procedures for communicating GAI incidents and performance to AI Actors and downstream stakeholders" sits with each individual company, rather than being tied to a broader, externally validated framework of likely actions defined against actor types. Given the ongoing uncertainty, it is worth considering how to provide more specific guidance in this area, while being cognizant of the need for companies to have flexibility in the face of commercial, safety and privacy concerns.

As noted above, participants in our survey reported that there needs to be greater transparency and more information passed from model providers to deployers, and vice versa. However, to be practical, this information should be of the most critical nature or "high impact disclosures", so as to avoid the sharing of superfluous or sensitive information which would increase the workload of downstream developers in reviewing and filtering all the information provided, and potentially expose companies to further risk. Such "high-impact" disclosures could involve, for example, any unexpected outputs from the model, changes to model performance, or risk management measures.

One interviewee from a small company described the current process of obtaining information about generative AI models as "ad hoc", and stated that they would welcome a more standardized process that was data protection compliant; did not risk the exposure of commercially sensitive information or intellectual property; and provided useful context for downstream developers without greatly increasing time spent on development. The challenge as articulated by participants was in determining exactly what information is critical or high-impact in each case, and ensuring that this is conveyed in a convenient and effective manner which takes into consideration the opportunity cost of preparing or engaging with this documentation. While the final version of the Generative AI Profile provides somewhat more guidance on managing risks associated with using third-party or pre-trained models, it could be further enhanced by explicitly clarifying the division of responsibilities between foundation model providers, hosts and deployers, and what information likely needs to be shared by each of these actors.

Systems and model cards could be part of the solution to this issue, and some model providers have started to release these alongside their AI models[9] where they have been well-received by the developer community. These initiatives should be considered an important step in the development of effective documentation which strikes the right balance between comprehensiveness and usability — for certain models and cases — but due to the complexity of assessing risk and concerns over inadvertent sharing of confidential data highlighted in the preceding paragraphs, it is presently very challenging to determine definitively what information a system or model card "should" include, and providers would need to maintain discretion over their design and deployment. Providing a range of formats and types of "cards" should be explored.

Table No. 3 — Potential information types to be included in model templates.

Drawing from industry practices and participant feedback, several categories of information emerge as potentially valuable for sharing (for example):	
Model documentation	Including model cards, intended use cases, and known limitations.
Risk assessments	Sharing insights on identified risks and mitigation strategies.
Performance metrics	Key indicators of model performance and reliability.
Incident reports	Information on unexpected behaviors or outputs, with clear criteria for what constitutes a reportable incident.
User feedback	Aggregated and anonymized insights from end-users, where appropriate.

Note: This non-comprehensive list of potential artifacts aims to capture the different types of content and transparency documentation which might be shared by different AI actors across the generative AI ecosystem.

From the point of view of the model providers, they also reported that it would be useful to have more information from downstream model optimizers and deployers. This information might include sharing planned or actual use cases, evaluations and red-teaming results, reports of adverse outcomes or unexpected behavior and other information about model performance in the deployment context which would enable them to anticipate and better guard against new risks and attack types. At present, according to respondents, it is particularly unclear what constitutes an "incident", what the criteria may be for reporting, and the best practices around this (see more below in Section 3.1).

DETAILS

Furthermore, participants suggested that deployers — who have the best access to end-user feedback and reporting — could share more of this direct feedback from users with model developers and providers, though there was concern expressed that sharing this kind of information directly from users might pose data privacy risks, and participants felt that they would need support from NIST or another body to help determine how to assess and mitigate this risk. We summarize the key insights from this section in the recommendation below.

☆ RECOMMENDATION

Expand and refine the AI value chain: actors, roles and responsibilities

NIST should conduct research pursuant to updating the taxonomy of AI Actors (perhaps in cooperation with the OECD), so that it better reflects the complex dynamics and relationships within the AI ecosystem. With these new actors and value chain model established, a new mapping of actions to actors could then provide more clarity for companies, and support their development of an efficient risk management system with clear roles and responsibilities. Additional guidance is needed on collaborative risk management between providers and deployers, with specific considerations for scenarios where roles are overlapping or ambiguous could be particularly valuable.

Additionally, NIST could consider facilitating a collaborative process with the AI community to develop a set of templates for information sharing up and down the AI value chain, or coordinating with and building upon the work of industry groups who have already done work in this area, such as Partnership on AI.[10] These templates should: (i) be flexible enough to accommodate different roles and use cases without overwhelming recipients or compromising sensitive information, (ii) prioritize "high impact" disclosures that are most critical for risk management; (iii) include clear guidelines on protecting sensitive information and intellectual property and align with existing data protection regulations; iv) acknowledge that there are some cases where it may not be possible or appropriate to share model information openly.

NIST could also explore the creation of a voluntary pilot program, “sandbox” or other research-focused program where different actors in the AI value chain can test and refine these information-sharing practices with real use cases in a safe environment. This approach would allow for iterative improvements based on real-world experience and provide actual evidence of the efficacy of information shared via the templates. The ultimate goal of the research would be to generate guidance and a set of flexible methods for companies who are developing and deploying products so they may safely and legally provide further insights into end user experiences and the prevalence of various actualized risks to upstream parties. Methodology or technical guidance for anonymizing and reporting on end-user use patterns and experiences would help encourage sharing and protect end user anonymity, trade-secrets, or other confidential information.

Finally, NIST should ensure that future guidance, including, for example, revisions to the draft NIST AI 800-1, aligns with the comprehensive approach of NIST AI 600-1 by addressing the roles and responsibilities of all actors across the AI value chain. This consistency is crucial for effective risk management and reflects the ecosystem-wide nature of AI development and deployment.

2.2 Open-source AI guidance

The NIST Generative AI profile does not specifically address the spectrum of open-source AI development and release strategies possible for companies, though it does mention open-source in the context of value chain and component integration risks. This limited treatment of open-source approaches — which is mainly addressed within the section on "Third-party Considerations"[11] — was a significant concern raised in our expert focus groups, and in both sessions different experts affirmed that source development plays a crucial role in the AI ecosystem, supporting competition, consumer choice and innovation.

FINDING

2.2.1 Open-source and spectrum of adoption

DETAILS

Data gathered throughout the second phase of this program revealed a strong need for guidance specifically focused on open-source approaches to generative AI development and deployment, with 63% of respondents stating that such guidance would be "very" or "extremely" useful. This underscores the growing significance of open-source models in the AI landscape, and the need for risk management strategies tailored to this context as more individuals and organizations combine elements of data, models and code which are available and licensable to varying degrees.

Within our cohort of participating companies the main reasons cited for using or creating fully or partially open-source models, data or tools were: cost-effectiveness, access to advanced technologies, and flexibility in customizing solutions. This view is consistent with a number of other experts and organizations who have asserted that open-source AI models and components accelerate scientific and commercial innovation[12], enhance transparency and help mitigate bias[13], enable independent researchers to identify and help fix design flaws, promote competition and reduce market concentration[14], and potentially combat inequality by providing broader access to AI capabilities.[15]

DETAILS

Yet a small number of companies on the program noted that along with the benefits, the decentralized nature of open-source spectrum development may bring some unique challenges if they are required to track and account for the contributions of other actors in the chain. Transparency, accountability and collaboration were seen as crucial for effective risk management across the open-source ecosystem and this may require some more nuanced approaches to risk calculation and distribution —indeed, one of the experts in our focus groups noted that within the draft Generative AI Profile there is an assumption of a "single central actor who will undertake all risk identification, mitigation and documentation work", when in reality open-source development is executed by multiple parties, and that this is foundational to ensuring the benefits are equitably realized. This expert cited the measures GV-1.4-001 and GV-1.4-002 as measures which would necessarily involve more than one actor in the value chain if they are to be successfully executed, and suggested that NIST may consider taking this into account in developing their tools regarding the AI value chain, with a broader spectrum of open-source development approaches receiving specific mention.

As mentioned in the preceding section, the U.S. National Telecommunications and Information Administration (NTIA) issued a report this summer entitled "Risks and Benefits of Dual Use Foundation Models". This report found that currently, the benefits of releasing model weights and making them available outweighed the potential risks. Among other benefits outlined in the report, NTIA found that open-weight models "diversify and expand" the array of actors involved in model development, decentralizing market control; allow developers to build upon and adapt previous work, broadening the availability of AI tools to small companies, researchers, nonprofits, and individuals and increasing confidentiality and data protection."[16] In the report NTIA did also note that: "at the time of this Report, current evidence is not sufficient to definitively determine either that restrictions on such openweight models are warranted, or that restrictions will never be appropriate in the future".[17]

FINDING

2.2.2 Opportunities for enhancing collaboration and transparency across the AI value chain

DETAILS

While the call for additional guidance on “open-development” approaches was clear among companies, there were some nuances in their requests. SMEs and startups specifically emphasized the need for simplified, resource-efficient approaches to managing the most severe risks such as the generation of material relating to biochemical weapons or child sexual abuse material (CSAM), often seeking ways to leverage community resources to enhance their capabilities. In contrast, larger enterprises who typically had already developed advanced systems for managing severe risks grappled more often with aligning open-development AI practices with existing corporate governance structures. These companies were more focused on mitigating reputational and compliance risks which may arise from the complexity of implementing multiple frameworks — this was especially true of those operating in highly regulated industries such as finance and healthcare.

We acknowledge that the current version of NIST’s Generative AI Profile has made improvements in addressing open-source considerations, specifically in the "Value Chain and Component Integration" risk category. Additionally, actions across several different AI RMF functions seem to address open-source considerations. These improvements indicate NIST’s acknowledgement of the importance of open-source in the AI ecosystem and largely align with the feedback received from participants. For example, the final version of the Profile now includes actions such as:

DETAILS

Table No. 4 - Overview of Open-Source AI Considerations in NIST's Generative AI Risk Management Framework 600 (July 2024 version).

Action ID	Explicit Mention of OS	Implicit Mention of OS	Description
GV-1.6-002		✓	Inventory exemptions for embedded Generative AI systems
GV-1.7-002	✓		Factors for decommissioning, including open-source data or models
GV-4.1-003		✓	Oversight across Generative AI lifecycle, including supply chains
GV-6.1-009	✓		Due diligence for open-source Generative AI tools
GV-6.1-010	✓		Acceptable use policies for open-source Generative AI technologies
GV-6.2-002	✓		Documenting incidents involving open-source software
MP-4.1-007		✓	Re-evaluating fine-tuned third-party models
MP-4.1-008		✓	Re-evaluating risks when adapting models to new domains
MS-1.1-001		✓	TEVV practices for third-party systems
MS-2.5-005		✓	Verifying training data provenance and grounding
MG-1.3-001		✓	Considering different approaches for model release
MG-2.2-009		✓	Responsible use of synthetic data techniques
MG-2.4-001	✓		Deactivation process for open-source models
MG-3.1-001		✓	Applying risk controls to third-party Generative AI resources
MG-3.1-005		✓	Reviewing transparency artifacts for third-party models
MG-3.2-002		✓	Documenting adaptations of pre-trained models
MG-3.2-003		✓	Documenting sources and types of training data

Note: This selection of suggested actions focus on various aspects, including due diligence for open-source tools, risk management practices specific to open-source integration, and considerations for third-party models that may leverage open-source components. While there are a number of explicit mentions, many more actions could have implicit relevance to open-source AI.

DETAILS

The growing adoption of open-source AI presents both opportunities and some challenges for organizations. While the NIST AI RMF 1.0 and Generative AI profile provide a strong foundation for AI risk management, there is room for improvement in addressing the unique aspects of open-source AI, and for adaptation to new capabilities as the technology develops. NIST can empower all actors in the AI value chain to more confidently and securely leverage open-source AI tools and models by expanding guidance in the areas identified and facilitating the collaborative development of best practices around open-source AI development and risk management. Further, experts also indicated the need for proportional requirements based on model capabilities and risks, as well as the importance of transparency in open-source AI development.

Future NIST AI-related guidance should also address these gaps. For example, as of the time of writing, the current approach taken by draft NIST AI 800-1 places responsibilities on open-source developers that do not take into consideration the context in which models are developed and deployed. This is both at odds with the Seoul Frontier AI Safety Commitments, and fundamentally incompatible with open-source development.
The draft guidance suggests measures such as the ability to decommission released models, which is particularly challenging in an open-source context where models can be freely distributed and modified. This approach might not fully align with the nuanced perspective presented in the NTIA report, which advocates for a Marginal Risk Framework. The report states: “The consideration of marginal risk is useful to avoid targeting dual-use foundation models with widely available weights with restrictions that are unduly stricter than alternative systems that pose a similar balance of benefits and risks”.[18] Adopting this framework would allow for a more balanced assessment of open-source models, considering both their unique risks and their substantial benefits to US interests, innovation and research.

☆ RECOMMENDATION

Provide more detailed guidance specific to the variety of open-source and full spectrum approaches and contexts

NIST should produce a detailed set of suggested actions tailored to open-source AI contexts, guidelines on secure data handling and model fine-tuning in open-source contexts, and strategies for managing the rapid update cycles of open-source AI models and components.

NIST should develop a simplified open-source AI adoption framework for smaller companies (SMEs and startups) — this should provide streamlined guidance for smaller companies to adopt open-source AI responsibly, including a "quick start" guide for key risk assessment and mitigation tailored to resource constraints, templates and checklists for basic risk management, and a roadmap for scaling practices as organizations grow. Similar to what NIST has already done in the context of cybersecurity through its “Small Business Quick-Start Guide”.[19]

Finally, NIST should consider implementing a marginal risk framework when assessing open-source foundation models, similar to the approach outlined in the NTIA report. This framework would compare the incremental benefits and risks of open-source AI technologies against existing technologies, providing a more balanced and context-aware assessment.

2.3 Interoperability of AI risk management frameworks

Phase 2 further revealed significant challenges faced by companies in navigating the complex landscape of AI governance frameworks. A high number of participating companies, particularly those operating internationally, reported struggling with the multiplicity and overlap of frameworks, with 47% stating that they are challenged by implementing frameworks across different jurisdictions.

Multinational corporations reported grappling with on average three different AI risk management frameworks across their global operations, leading to increased compliance costs and operational complexities. The three most commonly cited frameworks were the NIST AI RMF 1.0, the EU AI Act and the ISO/IEC 23894:2023[20]. As noted by one company — *“We're dealing with a patchwork of guidelines, standards and regulations. It's incredibly resource-intensive to ensure we're compliant across all jurisdictions."*

This challenge is exacerbated by the lack of harmonization or interconnections between different frameworks, creating uncertainty about which standards take precedence in cases of conflict, or what actions could adversely affect compliance or validation in other jurisdictions.

Our group of experts further emphasized the need for better alignment and interoperability between various AI governance frameworks. They echoed what organizations mentioned about their challenges with navigating multiple frameworks. One expert highlighted the challenge of “framework fatigue”. They emphasized the need for a clear “alignment map” showing how these different standards interact and complement each other, and suggested that NIST should be the central agent to “bring it all together” and create a single, centralized framework. This points to the need for NIST to not only align with other frameworks but also to provide guidance on how organizations can efficiently navigate the broader regulatory landscape.

NIST has recognized these challenges and is taking proactive steps to address them, as outlined in its AI 100-5 document, "A Plan for Global Engagement on AI Standards." The document emphasizes promoting global alignment, enhancing stakeholder engagement, and grounding standards development in scientific principles. These efforts provide a foundation for improving framework interoperability and accessibility.

NIST AI 100-5 - A Plan for Global Engagement on AI Standards[21]

NIST's AI 100-5 publication, "A Plan for Global Engagement on AI Standards," provides a broader context for addressing interoperability and harmonization challenges. Key points include:

1. Emphasis on developing science-backed, consensus-based standards through open, transparent processes.
2. Recognition of the need for standards that reflect diverse global stakeholder needs.
3. Prioritization of engagement in standards work, including pre-standardization research.
4. Commitment to facilitating diverse multi stakeholder participation in AI standards development.
5. Promotion of global alignment on AI standards approaches.

The publication identifies three categories of AI standards: those that are urgently needed and ready for standardization — such as standards on AI concepts and terminology, testing, evaluation, verification, and validation — and AI risk management tailored to specific contexts and risks. The second category includes standards that are needed but require more scientific work before standardization, such as conformity assessments. The third category consists of standards that are needed but require significant foundational work, such as techniques for measuring interpretability and explainability of AI system outputs.

This document can serve as a guide for the agendas of AI Safety Institutes being established in various countries, such as the US, UK, Canada, Japan, Singapore, and The Republic of Korea, and can also inform bilateral dialogues like the US-EU Transatlantic Technology Council.

The findings underscore the critical need for enhanced interoperability and harmonization of AI governance frameworks, especially in the context of generative AI risk management. The complexity and rapid evolution of generative AI technologies demand a more agile, coordinated approach to governance that can keep pace with technological advancements while providing clear, actionable guidance to organizations of all sizes.

☆ RECOMMENDATION

Further develop an interactive tool or “alignment map” which would map principles and practices across the most used generative AI risk management frameworks.

NIST should continue its efforts in developing comprehensive “crosswalks” between major AI governance frameworks, coupled with interactive tools to help companies navigate requirements based on their specific context and sector. For example, NIST in collaboration with other international partners could establish an online, interactive tool for companies to navigate framework requirements based on their location, generative AI use cases, development practices and other dimensions.

NIST should specifically consider aligning its risk taxonomy more closely with international frameworks, such as the OECD's work on AI incidents and hazards. This would promote global interoperability in AI risk management and incident reporting, facilitating more effective cross-border collaboration and knowledge sharing.

THE ROLE OF AI SAFETY INSTITUTES IN PROMOTING INTEROPERABILITY

The group of experts highlighted the establishment of the U.S. AI Safety Institute (AISI), which is part of NIST, as offering another significant opportunity to further address some of these challenges. The AISI’s mandate to evaluate AI systems, conduct fundamental safety research, and facilitate information exchange, positions it well to lead international efforts in developing globally recognized AI safety standards. This initiative could play a crucial role in bridging the gap between high-level principles and practical implementation guidance, particularly for generative AI technologies where benchmarks for LLMs and standards need to be set.

Regarding AI Safety Institutes, experts saw potential for these organizations to play a crucial role in advancing AI governance and risk management. However, they also raised some concerns and suggestions:

- Potential for coordination: experts suggested that newly created AI safety institutes could be a good forum for discussing and refining AI governance frameworks.
- Need for diverse expertise: there was emphasis on ensuring that AI safety institutes incorporate a wide range of expertise, not just technical but also from social sciences, ethics, and policy domains.
- Concerns balancing openness, expertise and inclusivity: Some experts raised questions about the composition and selection process for these institutes, stressing the importance of diverse representation. Experts also discussed how to balance the need for open, inclusive processes with the need for specialized expertise in developing and refining frameworks.

2.4 Evaluating the effectiveness of the NIST AI RMF

While our research has provided valuable insights into the implementation and practical challenges of the NIST AI RMF 1.0, it also highlights the need for ongoing evaluation of the framework's effectiveness. As noted on page 19 of the RMF document: "Evaluations of AI RMF effectiveness —including ways to measure bottom-line improvements in the trustworthiness of AI systems — will be part of future NIST activities, in conjunction with the AI community."

We recommend that NIST consider establishing a dedicated workstream to co-design and explore methodologies for assessing the AI RMF's effectiveness with the broader AI ecosystem. This effort could build upon exercises like the Open Loop program, which provide empirical data on the framework's real-world application.

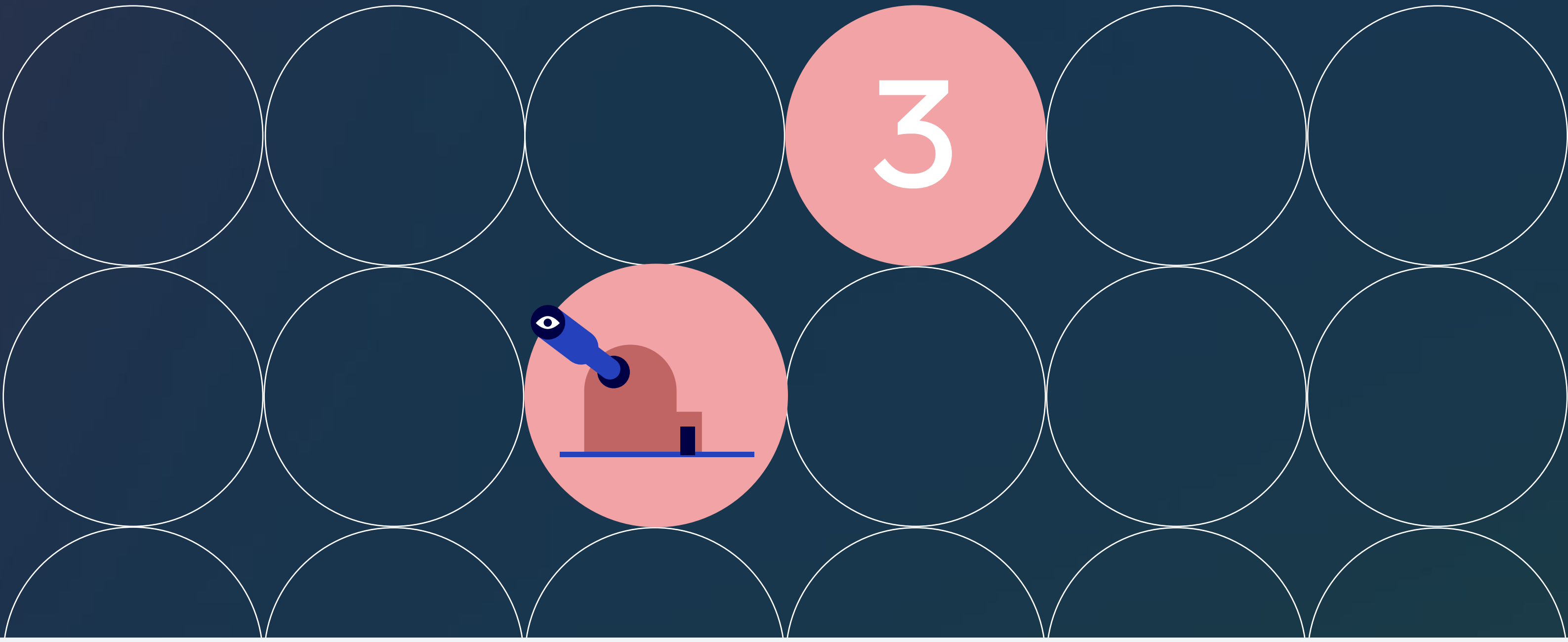
☆ RECOMMENDATION

Consider establishing a dedicated workstream to co-design and explore methodologies for assessing the RMF's effectiveness with the broader AI ecosystem.

This effort could build upon exercises like the Open Loop program, which provide empirical data on the framework's real-world application. Such a workstream could develop metrics to measure the framework's impact more holistically, establish mechanisms for regular user feedback, conduct longitudinal studies, collaborate on independent evaluations, and create a public repository of case studies and best practices. NIST can ensure that the framework continues to evolve in response to the rapidly changing AI landscape and the needs of its users. This approach aligns with NIST's commitment to maintaining the AI RMF as a "living document" and would provide valuable empirical evidence to inform future updates and refinements. In doing so, NIST can ensure that it remains a relevant and impactful tool for managing AI risks across diverse contexts and applications.

CHAPTER THREE

Findings & policy recommendations specific to NIST Generative AI Profile 600-1



Building on our Phase 1 report recommendations, which called for a more comprehensive risk taxonomy for generative AI,[22] this first part of this chapter examines the approach taken in NIST's AI 600-1 Generative AI Profile to categorizing and describing AI risks. In the second half, we discuss the opportunities for ensuring that the guidance is optimized for effective use.

3.1 List of generative AI risks

The Generative AI Profile introduces 12 key risks unique to or exacerbated by generative AI and provides a set of actions to help organizations govern, map, measure, and manage these risks across the AI lifecycle. While this taxonomy represents a significant step forward, our research indicates that further refinement could enhance its clarity, usability, and effectiveness.

List of generative AI risks

- CBRN Information or Capabilities
- Confabulation
- Dangerous, Violent, or Hateful Content
- Data Privacy
- Environmental Impacts
- Harmful Bias or Homogenization
- Human-AI Configuration
- Information Integrity
- Information Security
- Intellectual Property
- Obscene, Degrading, and/or Abusive Content

Our survey results show that 64% of participating companies found the 12 generative AI risks defined in the Generative AI Profile to be comprehensive, and did not suggest any additions to this list. Among the remaining 36%, there were a variety of suggestions as to how the list of risks could be supplemented or indeed approached in a fundamentally different way. For example, one participant offered a different way to categorize the risks — *"[I would] recommend splitting the list into two categories: "output risks'' and "model risks." Splitting the risk list in this way would better match the product development cycle, ensuring all actions have a risk associated, and providing specific scenarios or adverse events associated with each risk."*

Other suggestions included taking a more holistic approach which goes beyond a focus on risks and aligns approach more closely with the original concept of “Trustworthy AI” presented in the AI RMF 1.0 Playbook — *“I would change the approach. The original NIST AI document had 7 characteristics of trustworthy AI systems. Rather than focus everything on risks, focus on what people have to do to make the systems worthy of our trust”.*

Organizations across different sectors and of varying sizes found it difficult to adapt the generic risk descriptions to their specific use cases. This challenge is particularly acute for emerging AI applications and in rapidly evolving technological landscapes, where the relevance and magnitude of certain risks may shift quickly. There was more demand to further connect this taxonomy with other established sector-specific risk taxonomies which could provide more insight into how risks are likely to manifest in industries such as healthcare or finance.

Expert focus groups emphasized the need for a more dynamic risk assessment approach. One expert noted, “*The current taxonomy is a good start, but we need to consider how these risks interact and potentially amplify each other in real-world scenarios.*” Many of the listed risks are potentially interrelated and may have compounding effects under certain conditions. By not articulating this possibility, the current framework might inadvertently encourage organizations to address risks in isolation, potentially limiting the effectiveness of their risk mitigation strategies. This suggests that future iterations of this list of risks could benefit from a more interconnected view of risks, possibly including a risk interaction matrix.

Experts also emphasized the importance of clearly defining and differentiating between risks, hazards, and harms – a distinction that is not explicitly made in the current framework. For example, recent work by the OECD on defining AI incidents and related terms provides valuable insights that could inform NIST's approach. The OECD has proposed distinctions between AI incidents (actual harm) and AI hazards (potential harm), as well as gradations of severity within these categories. This framework offers a more nuanced understanding of AI-related risks and their impacts.[23]

With regards to specific risks listed, we also gathered from participating companies and expert discussions the following:

- The term "bias" was seen as potentially misleading, as it encompasses a wide range of concerns that might be better addressed separately;
- An additional risk category that emerged from our research, particularly from feedback provided by companies, relates to employee misuse and lack of control in the workplace. Two participating companies highlighted the risk of employees accessing and potentially misusing generative AI tools without proper oversight; effectively creating an invisible value chain risk. Traditional risk management approaches in the buyer-supplier value chain may not apply in these cases. All organizations expressed the need for better tools and policies to manage employee access and ensure compliance, similar to insider threats in cybersecurity.

It is commendable that the final version of the Generative AI Profile issued by NIST in July 2024 has addressed this issue by providing more detailed and nuanced descriptions for each risk category. For example, the "Confabulation" risk now includes a more comprehensive explanation of how this phenomenon occurs and its potential impacts. This improvement helps organizations better understand the nature and implications of each risk. However, there remains room for expanding upon these descriptions in future versions.

Future NIST AI-related guidance should also address these gaps. For example, as of the time of writing, the current approach taken in the draft NIST AI 800-1 guidance seems to present a list of risks that foundation model developers should address without adequately contextualizing them within a Marginal Risk Framework or differentiating how these risks might vary across different stages of the AI lifecycle and various deployment contexts, as recommended in the NTIA report. This approach allows for a more nuanced understanding of the incremental risks posed by dual-use foundation models compared to existing technologies.

RECOMMENDATION

Adapt the list of risks into a dynamic, inter-relational risk matrix with sub-profiles for specific sectors and more detailed descriptions of the risks outlined

NIST should iterate on this list and create a relational framework that acknowledges the interconnected nature of risks, showing how they interact and potentially compound each other. This could involve a matrix approach or a hierarchical system that groups the 12 risks into broader categories (e.g., output risks, model risks, operational risks) while maintaining detailed individual risk descriptions. For example, "Information Integrity" risks could be shown to have direct impacts on "Human-AI Configuration" risks.

NIST could further provide more detailed, sector-specific “sub-profiles” and real examples to help organizations translate generic risk descriptions into actionable insights for their particular domains. This could include guidance on risk prioritization within and across categories, helping organizations focus their risk management efforts more effectively, in most cases ensuring efficiencies and avoiding duplicative efforts. Moreover, to avoid placing infeasible requirements on foundation model developers, NIST should provide necessary technical guidance to enable all organizations to conduct structured, pragmatic risk assessments. This should include clear guidance on prioritizing the most critical risks to mitigate, helping companies allocate resources effectively.

NIST could also clarify the distinction between risks, hazards, and harms. Clear definitions and examples should be provided to help organizations differentiate between these concepts. This distinction is crucial for effective risk management. For example, bias in training data could be classified as a hazard; discriminatory outputs as a risk; and negative impacts on marginalized communities as a harm.

NIST could further integrate guidance on anticipating and preparing for emerging and long-term risks. This could involve scenario planning exercises and regular updates to the risk taxonomy to reflect the rapidly evolving AI landscape.

Finally, future NIST guidance should strike a balance between comprehensive risk management and practical implementation, considering the diverse capabilities and resources of different AI actors across the value chain. In the context of emerging AI guidance (such as proposed by draft NIST AI 800-1) NIST could revise its risk taxonomy to incorporate a Marginal Risk Framework, as recommended in the NTIA report. This should include, for example: (i) clearly differentiating between risks that are unique to or exacerbated by dual-use foundation models versus those common to other AI or non-AI technologies, (ii) providing guidance on how to assess the incremental impact of these risks in various deployment contexts and (iii) offering a methodology for prioritizing risks based on their marginal impact and likelihood in specific use cases.

3.2 Usability and practical implementation of the guidance

While the majority of the respondents are planning to use NIST's Generative AI Profile in the next 6-12 months, only 7% had already operationalized it by the time our data collection concluded in July 2024. As many as 40% reported that they are not planning to use the guidance in the "near future".

The survey also revealed that 90% of respondents expect implementing the Generative AI Profile to create either a "high workload" (45%) or a "very high workload" (45%), with its over 400 "suggested actions". Participating companies also identified the following organizational factors as the most salient obstacles to implementing the guidelines: limited staffing (73%), complexity of the guidance (55%), lack of financial resources (36%), and insufficient technical expertise (36%).

Companies, particularly smaller ones, expressed concerns about the practicality of implementing the guidance, specifically given the comprehensive nature of the actions covered, as encapsulated in this quote by one of the participants: "A document with hundreds of recommendations is not practical. We need a way to aggregate and simplify to accelerate adoption."

Furthermore, the companies in the program felt that there were still improvements to be made on the clarity and specificity of the language used in the guidance. For example one of the interviewees noted: "It is quite vague — how should 'Conduct regular audits of third-party entities to ensure compliance with contractual agreements' be done? What is the recommended frequency? Figuring this out may be very easy for a big company, but it will be very challenging for SMEs". Overall, organizations highlighted somewhat similar resource related challenges in implementing the NIST Generative AI Profile guidance as indicated during Phase 1 Open Loop report when asked about the AI RMF 1.0.

Survey results and interviews revealed a strong desire among organizations for practical resources such as templates, case studies, and compliance checklists to support implementation. The diversity in our cohort's size, maturity, and position in the AI value chain underscored the importance of creating guidance that is accessible and applicable to organizations at different levels of maturity with respect to AI development and deployment. Particularly, the challenges faced by small enterprises (39% of our cohort) in implementing comprehensive responsible AI practices highlight the need for scalable and resource-efficient approaches to AI risk management.

For example, 64% of survey respondents found "templates for incident response and reporting" extremely useful,[24] and 45% found "case studies demonstrating successful risk management practices" extremely useful. One expert suggested, *"We need resources that can evolve as quickly as technology does. Static checklists won't cut it in the world of generative AI."* This indicates a need for NIST to consider developing dynamic, possibly AI-assisted tools that can keep pace with the rapid evolution of generative AI technologies and their associated risks.

The final version of the NIST Generative AI Profile has made significant strides addressing most of these concerns. The "Suggested Actions to Manage Generative AI Risks" section now includes more executable guidance, with each action tagged with relevant “AI Actor Tasks”.[25] The revised version now has around 200 suggested actions instead of around 400 (draft version). It is yet to be seen how this improvement could potentially help organizations identify which actions are most relevant to their role in the AI ecosystem, potentially reducing the perceived workload. However, in order to make this additional guidance even more useful, NIST should consider providing a chart which is broken down by Actor, rather than action. In doing this they would greatly increase the usability and legibility of the guidance, while creating a non-exhaustive, non-exclusive list of actions listed by Actor for easy reference.

While these enhancements are substantial and address many of the concerns raised by participants, there are still opportunities for further improvement.

☆ RECOMMENDATION

Consider the usability (or UX) elements of the Generative AI Profile and companion guidances, focusing on enabling small companies to efficiently find and understand their responsibilities

NIST could develop a tiered guidance system that caters to organizations of different sizes and AI maturity levels. This could include a "quick start" or reader guide for smaller organizations or those new to AI risk management. Over time, NIST could provide more case studies and examples of how organizations have implemented the profile’s suggested actions that would make the guidance more concrete and easier to apply. Further, NIST could create online tools and video explainers to guide users through the risk assessment process, helping them identify which parts of the profile are most relevant to their specific context, or even a “self-assessment” simulator for organizations to measure their compliance/adoption levels.

CHAPTER FOUR

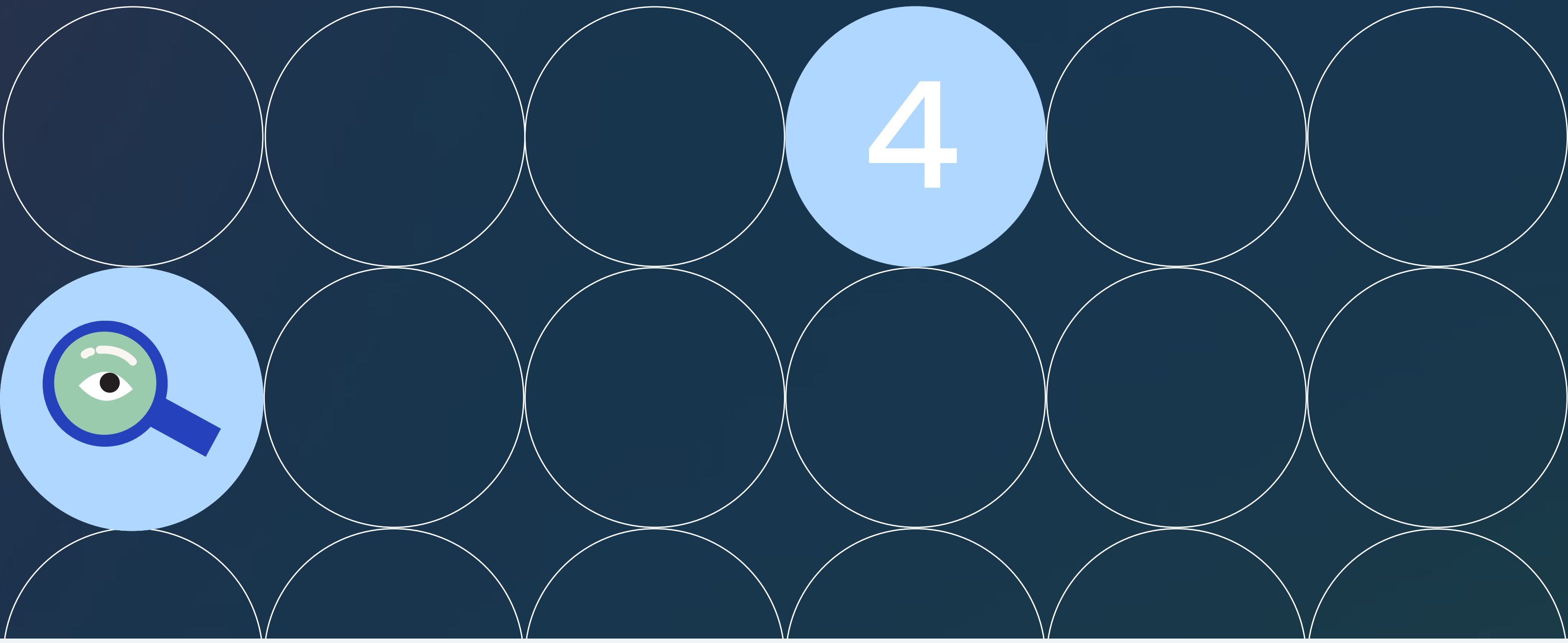
Conclusion

Over the past year, the Open Loop US program on Generative AI Risk Management has provided an analysis of the challenges and opportunities in implementing effective AI risk management practices, with a particular focus on the work of NIST in regard to the AI RMF 1.0 and the Generative AI Profile.

Our Phase 1 report highlighted key opportunities for supporting organization's in their journey to responsible generative AI development and deployment, including the need for a comprehensive, generative AI-specific risk taxonomy, access to tools, and crucially — guidance on AI red-teaming and synthetic content risk management. It is important to acknowledge that NIST has made significant strides in addressing many of these areas with the release of the Generative AI Profile 600-1 and other documents issued in July 2024.

However, our Phase 2 findings revealed that while progress has been made, several gaps remain. Organizations continue to invest a significant amount of capacity in implementing robust AI risk management practices, particularly in the context of generative AI. The revised NIST guidance, while more comprehensive and targeted, still presents challenges in terms of practical implementation, especially for smaller organizations or those with limited resources.

As we look to the future, it's evident that effective AI risk management — and particularly in the context of generative AI — will require ongoing collaboration between policymakers, industry leaders, SMEs, AI researchers, and society as a whole. Open Loop remains committed to facilitating this dialogue and providing evidence-based insights to support the development of effective, practical AI governance frameworks.



References

[1] White House. (2023, October 30). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Retrieved from https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[2] NIST guidance including: "Reducing Risks Posed by Synthetic Content" https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf ; "A Plan for Global Engagement on AI Standards" https://doi.org/10.6028/NIST.AI.100-5; "Secure Software Development Practices for Generative AI and Dual-Use Foundation Models" https://doi.org/10.6028/NIST.AI.800-1.ipd ; "AI Risk Management Framework: Generative Artificial Intelligence Profile" https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

[3] National Institute of Standards and Technology. (2024, July 29). NIST Announces New Draft Guidance. Retrieved from: https://www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following

[4] NIST, Trustworthy and Responsible AI Resource Center. Retrieved from: https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/2-sec-audience

[5] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), page 8, Section 1.2.4. Retrieved from: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[6] Partnership on AI, "Risk Mitigation Strategies for the Open Foundation Model Value Chain", (July 2024), Retrieved from: https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

[7] NIST, "Generative AI Profile", NIST.AI.600-1. Retrieved from https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[8] See Report 1 of this Open Loop US program: https://openloop.org/reports/2024/01/red-teaming-synthetic-content.pdf

[9] Meta AI. (2022, February 23). System Cards, a new resource for understanding how AI systems work. Retrieved from https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/

[10] Partnership on AI. (2024, June). How Better AI Documentation Practices Foster Transparency in Organizations. Retrieved from https://partnershiponai.org/wp-content/uploads/2024/06/PAI_ABOUT-ML-pilots-summary.pdf

[11] NIST. (2024, July). Generative Artificial Intelligence Profile. Pag. 48. Retrieved from https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[12] Vanschoren, J. (2023). Democratising artificial intelligence to accelerate scientific discovery. In OECD, Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research. OECD Publishing. Retrieved from https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-science_be9632d7-en

[13] World Economic Forum. (2023, December). Why open-source is crucial for responsible AI development. Retrieved from https://www.weforum.org/agenda/2023/12/ai-regulation-open-source/

[14] European Commission. (2020, October). The impact of open-source Software and Hardware on technological independence, competitiveness and innovation in the EU economy. Retrieved from https://ec.europa.eu/newsroom/dae/redirection/document/79021

[15] Bateman et al. (2024). Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation Model Governance. Carnegie Endowment for International Peace. Retrieved from https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance&ved=2ahUKEwi6vb_cqc2IAxX80AIHHYw_HdcQFnoECBcQAQ&usg=AOvVaw2YtKoK2alt80ubAGsqNo53

[16] National Telecommunications and Information Administration (NTIA). (2024). Dual-use foundation models with widely available model weights. Retrieved from https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report

References

[17] Ibid, page 3.

[18] Ibid, page 10

[19] NIST, NIST Cybersecurity Framework 2.0: Small Business Quick Guide. Retrieved from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1300.pdf

[20] International Organization for Standardization. (2023). Information technology — Artificial intelligence — Guidance on risk management (Edition 1). Retrieved from https://www.iso.org/standard/77304.html

[21] National Institute for Standards and Technology (NIST), A Plan for Global Engagement on AI Standards. (2024, July): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5.pdf

[22] See Report 1 of this Open Loop US program: https://openloop.org/reports/2024/01/red-teaming-synthetic-content.pdf

[23] Organisation for Economic Co-operation and Development. (2024). Defining AI Incidents and Related Terms. Retrieved from https://www.oecd.org/en/publications/defining-ai-incidents-and-related-terms_d1a8d965-en.html

[24] The updated NIST Generative AI Profile now provides more detailed and specific steps in the "Suggested Actions" section. For instance, action GV-4.3-002 now includes specific criteria for Generative AI system incident reporting.

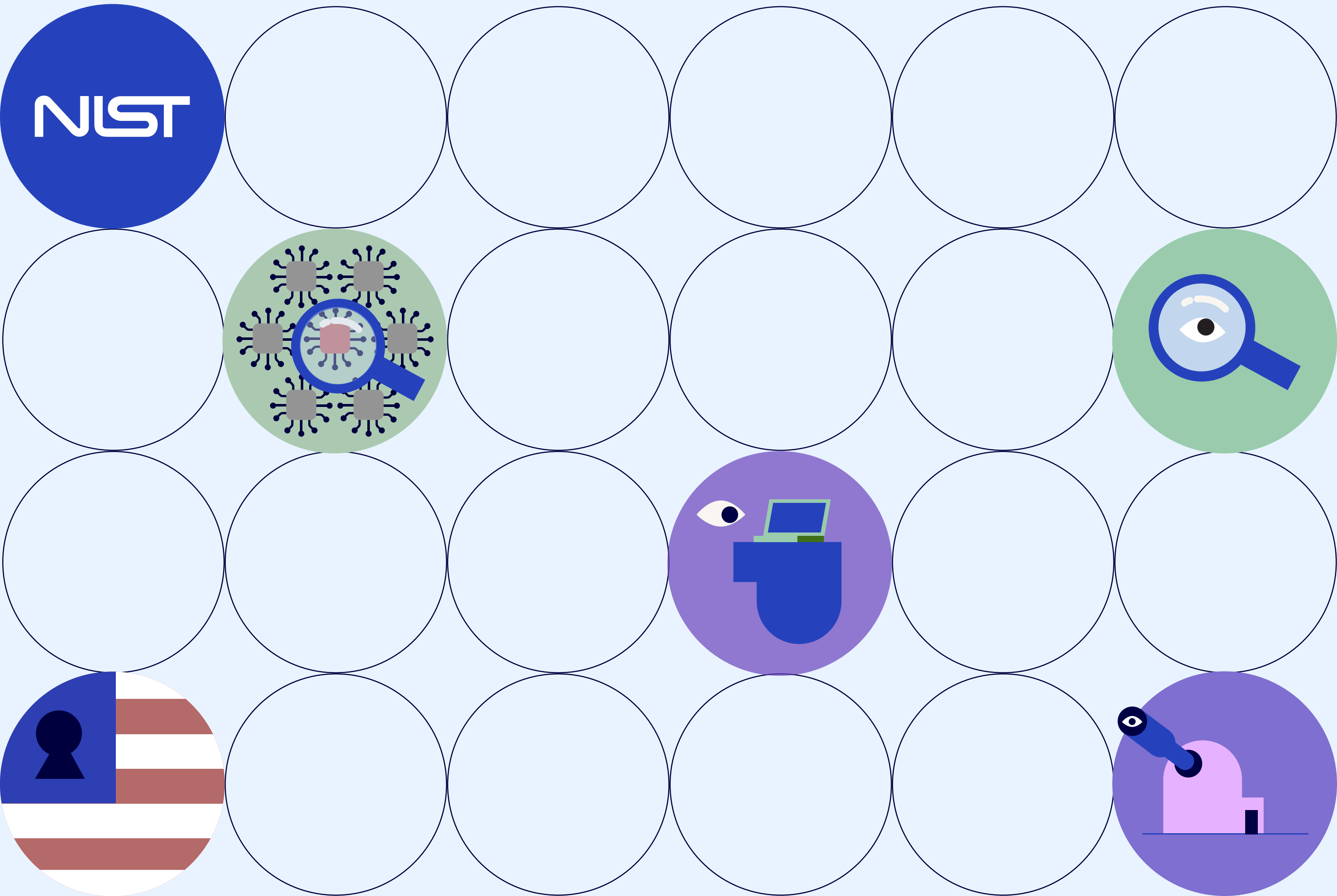
[25] See observations above related to the AI value chain and AI actors (Chapter 2, Section 2.1).

Program report 2

Open Loop | Meta

Generative AI Risk Management and the NIST Generative AI Profile (NIST AI 600-1)

Annex



Methodological Note: Open Loop US Program Phase 2 - NIST AI RMF 1.0 Adoption and draft Generative AI Profile Feedback

This methodological note details the research approach used in Phase 2 of the Open Loop US Program. This phase aimed to analyze the implementation of the NIST AI Risk Management Framework (RMF) and gather feedback on NIST's AI 600-1, the draft NIST Generative AI Profile. The program generated valuable insights into the experiences of organization's as they sought to implement this guidance. The findings were used to develop recommendations for improving the clarity, effectiveness, feasibility and usability of the NIST AI RMF, related AI guidance and the Generative AI Profile in particular, ultimately promoting the safe, secure and trustworthy development and deployment of generative AI technologies.

Goals

- Analyze Responsible AI (RAI) maturity and its influence on NIST AI RMF adoption.
- Understand company needs regarding the Generative AI Profile across different value chain positions and company sizes.
- Identify challenges in implementing the Generative AI Profile's risk categories and actions.
- Develop recommendations for improving the Generative AI Profile's clarity, effectiveness, and usability.
- Explore best practices for integrating NIST frameworks into existing risk management processes.

High-Level Research Questions

- How do organizations' current risk management practices and RAI maturity influence their adoption and integration of the NIST AI RMF?
- How do organizational needs regarding the Generative AI Profile vary based on their position in the generative AI value chain and company size?
- What specific challenges do companies face in understanding and implementing the risk categories and actions outlined in the Generative AI Profile?
- What recommendations can be made to improve the clarity, effectiveness, and usability of the Generative AI Profile?
- What best practices can be identified for integrating the NIST AI RMF and the Generative AI Profile into existing corporate risk management processes?

Sampling and focus

While Phase 1 involved a larger sample size of 40 companies to establish a broader landscape view which allowed us to identify trends and patterns across a larger group of companies, Phase 2 strategically adopted a more focused approach. In Phase 2, we wanted to gather specific, detailed feedback on the draft Generative AI Profile which could inform NIST's future iterations of this document.

To achieve this, we invited a subset of selected companies who had confirmed that they had analyzed the Generative AI Profile to take part in semi-structured interviews, prioritizing a more substantive understanding over a larger, potentially less detailed dataset. This selection process ensured the interviews would yield richer data, allowing us to delve deeper into how companies navigate the complexities of the Generative AI Profile and its implementation.

We recognized the importance of capturing diverse perspectives within the Generative AI ecosystem. Companies were chosen to represent various positions in the Generative AI value chain. This included not just large-scale entities, but also organizations acting as Generative AI model developers, deployers, and acquirers. The experiences of companies across this spectrum were aimed at a richer understanding of how the Generative AI Profile resonates with different stakeholders in the development and deployment of Generative AI technologies.

Cohort's profile

Phase 2's cohort represented a diverse cross-section of the AI ecosystem:

Value Chain Roles

67% downstream deployers, 47% GenAI-powered tool producers, 33% model acquirers.

Company Size

46% large enterprises (>250 employees), 39% small to micro enterprises (<50 employees) and 15% medium-sized enterprises (50 - 249 employees)

Company Age

77% established for over 5 years.

RAI Governance Maturity

40% rolling out RAI-specific practices, 27% formulating such practices.

This diversity provided insights from various organizational contexts and AI value chain positions, enriching our understanding of NIST AI RMF implementation challenges and opportunities.

Survey

- Online survey of 20 questions.
- Content: Quantitative data on risk management practices, RAI maturity, and NIST AI RMF awareness.
- Sample size: 16 respondents

We were able to leverage the insights from Phase 1 to conduct more focused and in-depth interviews in Phase 2. This two-phased approach, with its initial broad survey followed by focused interviews, ultimately yielded a comprehensive understanding of company needs and challenges regarding the NIST AI RMF and the draft Generative AI Profile.

Interviews

The interview structure in Phase 2 of the Open Loop US Program consisted of four key sections designed to gather in-depth insights into company experiences with the NIST AI RMF and the Generative AI Profile. The first section explored current AI governance practices, including roles and responsibilities for AI development, current policies related to AI principles, and data governance controls. Section two focused on the company's experience with the NIST AI RMF, delving into their integration processes, challenges encountered, and any alignment with existing risk management frameworks.

Number of interviews

14

Technical input through Expert Groups

In addition to company interviews, we also sought technical input through two focus groups conducted with our broader group of experts supporting the program design and testing. This approach complemented the insights gleaned from company interviews.

Focus Group 1 on Generative AI Risk List and Actions

The first expert group discussion centered on the list of risks and the complexity of actions outlined in the Generative AI Profile. Their technical expertise provided valuable feedback on the comprehensiveness and clarity of the risk categories, as well as the feasibility and practicality of the recommended actions.

Focus Group 2 on Open-Source Models and Generative AI Profile Coverage

The second expert group focused on the role of open-source AI foundation models within the Generative AI ecosystem and their coverage within the Generative AI Profile. Their insights helped assess whether the Profile adequately addressed the gaps in the Profile in addressing specific open-source AI considerations and areas where there could be further guidance for companies.

Limitations

Self-reported data

The self-reported nature of the questionnaire data may introduce some bias.

Limited sample size

The sample size for the semi-structured interviews was small, potentially limiting generalizability of the findings.

Location scope

The research focused on companies within the US, limiting insights into global trends and challenges, though many of the larger companies were also operating to some extent outside the US.

If you would like to know more about our data collection or analysis methods please reach out to us at usprogram@openloop.org

Open Loop | Meta

www.openloop.org