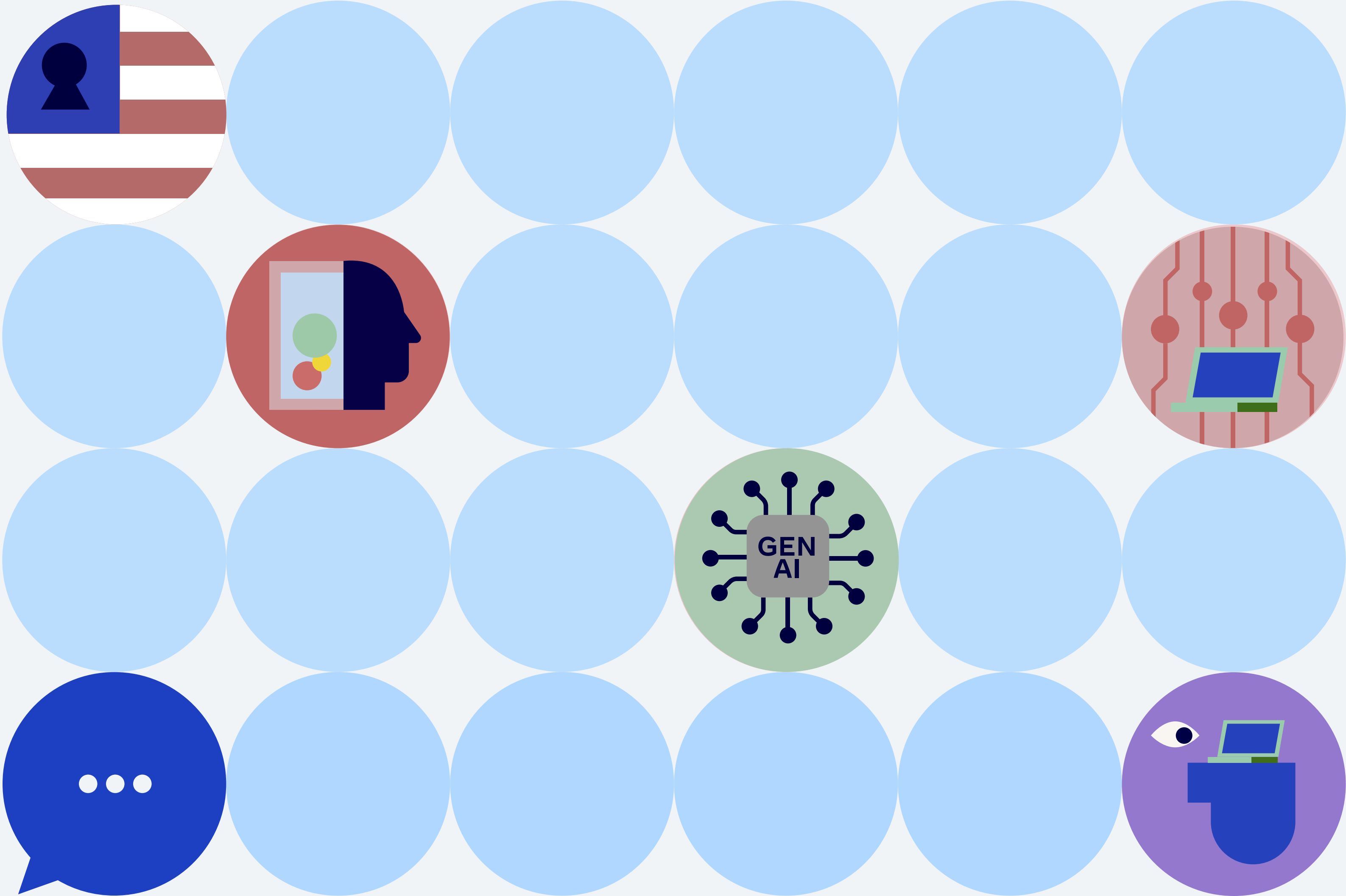
NIST AI Risk Management
Framework and Generative AI

Open Loop | Meta

Red-Teaming & Synthetic Content

This report presents the findings and recommendations of the first phase of the Open Loop US program on Generative AI Risk Management, focused around AI red-teaming and synthetic content risk mitigation.



Foreword

Generative AI is a catalyst for transformative changes in every industry and across the globe, and in order for us to fully realize the benefits of this technology, it needs to be understood and managed responsibly.

To harness its benefits while effectively managing the possible risks, developers and downstream deployers need clear, globally consistent risk management guidance, standards and benchmarks, especially for novel forms of AI. This guidance must be comprehensive, yet flexible enough to accommodate the fast pace of AI development and the needs of a wide array of users. It should be based on evidence of what works in practice for companies of all scales, and it must support AI innovation.

This Open Loop report on risk management and generative AI is a valuable input to the effort to collaboratively develop the next generation of AI risk management guidance and practices, and specifically to answering questions about how existing frameworks can be adapted and built upon to support the responsible deployment of generative AI systems.

The program takes as its testing subject the Artificial Intelligence Risk Management Framework 1.0 (AI RMF) and accompanying playbook and resources created by the National Institute of Standards and Technology (NIST). The program is split into two phases, and during the first phase we gathered feedback from 40 companies regarding their understanding of the AI RMF, their current risk management practices as they relate to red-teaming and synthetic content risk management, and where they see opportunities for enhancing the current NIST AI RMF 1.0 in these areas. The report shares our findings and recommendations. Our hope is that these will be useful to NIST as they discharge their duties over the coming months.

This work would not be possible without our partners in Accenture, the committed group of AI experts and thought leaders who have helped us shape the program, and the participating companies who have invested significant time and energy into this effort, my sincere thanks to all of them.

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

Foreword

At Accenture, we recognize the importance of responsible AI governance and the need to help ensure that the benefits of AI are realized while minimizing potential harm. That is why we are proud to support the National Institute of Standards and Technology (NIST) in their mission to provide guidance on the responsible use of AI.

As part of the Open Loop program, Meta and Accenture are partnering to bring together a wide variety of organizations and viewpoints to explore how to effectively govern generative AI by leveraging the NIST AI Risk Management Framework. This program has provided a platform for passionate advocates of responsible AI to share their insights and expertise, feeding into the work NIST is doing to develop the AI RMF and helping ensure that the development and use of generative AI aligns with societal values and expectations.

Closing the gap between responsible AI intention and action requires a commitment to move beyond frameworks and into practical plans. As we move forward, it is crucial that organizations embrace responsible AI governance and take concrete steps to implement it across their operations. We encourage organizations to adopt the NIST AI Risk Management Framework, establish AI governance structures and principles, conduct AI risk assessments, enable systematic responsible AI testing, and establish ongoing monitoring of AI systems. Together, we're building the next generation of trusted and secure AI technologies so that we may continue to harness the vast potential of AI for the benefit of society.

We would like to express our sincere gratitude to NIST, the AI experts and practitioners, start-ups, and enterprises who have generously contributed their knowledge and time to the Open Loop program, and to Meta for convening us all together and for their dedication, energy, and commitment to the safe and responsible use of AI.

Arnab Chakraborty
Chief Responsible AI Officer, Accenture

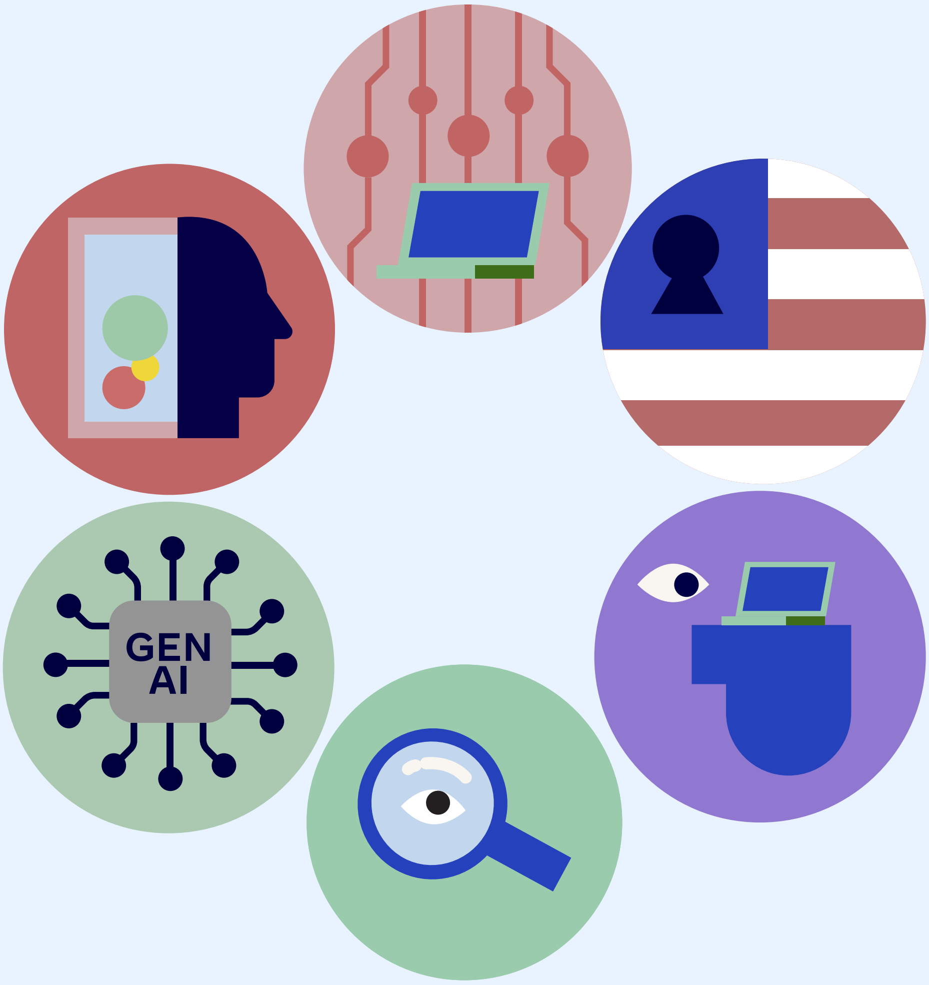
About Open Loop

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 a structured methodology, Open Loop participants co-create policy "prototypes" and test new or existing AI policies, regulations, laws, or voluntary frameworks. These multi-stakeholder efforts support rulemaking processes and improve the quality of guidance and regulations on emerging technologies, ensuring that they are understandable, effective and feasible in practice.

This report presents the findings and recommendations of the first phase of the Open Loop US program on Generative AI Risk Management, launched in November of 2023 in partnership with Accenture.

How to cite this report?
"Authors: Laura Galindo, Taja Naidoo, Maartje Nugteren, Ali Shah. Open Loop US program on Generative AI risk management: AI red teaming and synthetic content risk (2024)."



Acknowledgements

The Open Loop US program was created by Meta, and designed and implemented in collaboration with Accenture. We would like to thank the following organizations for their partnership and participation — without their commitment and active involvement, this project would not have been possible:

- 2021.AI
- Active Fence
- AI Collaborator Inc
- AI Squared
- Alt X Innovation AI Consultancy
- Amazon Web Services
- Applause App Quality Inc
- AT Worthy Technology, Inc.
- Avanade, incl.
- Beamery
- Best Buy
- Best Buy Health
- BlantonTechknowledge
- BluePrint Digital
- ChatBlackGPT
- Darktrace Federal
- DCI Solutions
- Direct Relief
- Holistic AI
- iJuno
- Inrupt
- IntelliDyne, LLC
- International Rescue Committee
- LegalMente AI
- Navy Federal Credit Union
- Numbers Protocol
- RAI Audit
- RevUp AI
- RWS Group
- Sky Hive
- StarDog
- Talespin
- Unifi.ai
- US Bank
- Velocity Global
- VU technologies Corp
- Zaviant

Additional companies elected to participate anonymously, including major multinational hospitality, consumer goods, and financial services brands. Their insights are included throughout.

Acknowledgements

Thank you 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 (CredoAI), Ellie Graeden (Luminos.Law; Georgetown University Massive Data Institute), Eugenio Zuccarelli, Evi Fuelle (CredoAI), Dr. Heather Frase (Georgetown's Center for Security and Emerging Technology), Jennifer Jin (Centre for Information Policy Leadership), JudeAnne Heath (Hispanic Technology & Telecommunications Partnership), 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 (Center for Growth & Opportunity, Utah State University), Nile Johnson (Institute for Security and Technology), Philip Dawson (Armilla AI), Dr. Rumman Chowdhury (Humane Intelligence), Stephanie Ifayemi (Partnership on AI), and Tom Romanoff (Bipartisan Policy Center).

We would also like to thank 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).

Thank you also to our design partners at Craig Walker Design and Research, in particular John-Henry Pajak.

Last but not least, thanks to the following colleagues at Accenture for their contributions to the program and this report — Lara Pesce Ares, James Baxani, Patrick Connolly, Tessa Darbyshire, Adeline Decuyper, Charlie Moskowitz, Rachel Pelham, Andrea Risberg, Haley Rosowsky, Manal Siddiqui, Shekhar Tewari, Dikshita Venkatesh, Jakub Wiatrak

Executive Summary

Open Loop is a global program that connects policymakers and technology companies to help develop effective and evidence-based policies around AI and other emerging technologies.

> The primary objective of this Open Loop Program on generative AI risk management is to assess the US National Institute of Science and Technology (NIST) AI Risk Management Framework (AI RMF) in practice.

This voluntary framework — released in January 2023 — aims to enable organizations to better manage AI risks to individuals, groups and society.

To facilitate the exploration of generative AI risk mitigation and the AI RMF more broadly, we designed this Open Loop in two phases. The first phase focused around two topics that are key to generative AI risk management and of particular interest for NIST, namely AI red-teaming and synthetic content risk mitigation. This report shares the results of the first phase of the program which took place from January to April 2024 and involved 40 companies. These companies represented a variety of industries and ranged in size from very large multinationals to medium-sized enterprises and startups

Through desk research, interviews, surveys and workshops, we investigated:

- ✓ How companies currently approach or plan AI red-teaming and/or synthetic content risk management efforts.
- ✓ The key challenges to efficient and successful implementation of AI red-teaming and/or synthetic content risk management.
- ✓ How the NIST AI RMF can be leveraged to resolve those challenges, enhance efficiencies, and support cross-value-chain collaboration.

Our findings indicate that:

→ Participating companies are motivated to adopt AI red-teaming and synthetic content transparency measures primarily to maintain customer trust, ensure regulatory compliance, and manage AI-related risks. These practices are seen as crucial for protecting brand reputation, navigating regulatory landscapes, and prioritizing resources effectively.

→ Both AI red-teaming and synthetic content transparency present several common challenges, particularly for small and medium-sized enterprises. Factors which exacerbate challenges are incomplete or unclear guidance, technical integration difficulties, resource constraints, and organizational and cultural barriers. These challenges form a complex web of interconnected issues, making them difficult to address in isolation, however with support these challenges are surmountable, and companies are making progress.

→ Among our participants there is a strong demand for comprehensive, clear and practical guidance from NIST, particularly on AI red-teaming and on identifying and mitigating risks related to synthetic content where these companies are currently unclear on both the strategies and tactics required for successful management of risks in these areas and how activities should be prioritized.

→ Open-source tools are seen by this group as crucial for lowering the barriers such as implementation costs and skills needed to stand-up new risk management systems or techniques, and for avoiding vendor lock-in.

From these findings, we have formulated the following recommendations:

1 Generative AI risk should be comprehensively identified and mapped.

For both red-teaming and synthetic content risk management, the key recommendation from the cohort is for NIST to develop a detailed taxonomy of specific generative AI risks to be addressed. NIST has taken the first step in addressing this need in their recent draft report, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.

2 Red-teaming guidelines could provide enhanced practical support.

In developing further red-teaming specific guidance, NIST should seek to support organizations in establishing the purpose and scope of generative AI red-teaming efforts, by providing a systems engineering framework for managing risk in AI systems (akin to NIST 800-160v1r1). That framework should articulate the appropriate use of detailed AI red teaming and scaled/automated measurement, and describe the basic steps that make up an assessment (similar to what is presented in NIST 800-30r1).

3 Enterprise-level metrics enable organizations to consistently assess quality and success in red-teaming efforts.

NIST should consider driving and supporting the setting of benchmarks by international standards organizations by gathering data on what the benchmarks should be for various red-teaming activities, and comparing and evaluating emerging benchmarks. As benchmarks are established and widely adopted, NIST may also be able to provide authoritative training data sets that can help organizations fine tune their systems or otherwise align with commonly accepted measurements.

4 Case studies and best practices on AI red-teaming are valuable.

Participants see collaboration and experience-sharing across the AI value chain as useful for fostering best practice and knowledge development within the AI ecosystem, and dedicated platforms or channels for information sharing should be explored to further encourage this activity.

5

Encourage the creation and provision of open source tools and techniques that enable efficient AI red-teaming.

NIST could encourage transparency regarding available tools, conducting a stocktaking exercise and classifying those which are currently available by capability and availability, with a view of which red-teaming activities a given tool would be most appropriate for.

6

Lead on the development of interoperable risk management frameworks

Frameworks and guidelines which do not interoperate are a barrier to the efficient scaling of risk operations, especially where companies are operating in multiple US states or internationally. NISTs recent plan for engagement on AI Standards is very welcome, and would be further enhanced by the addition of more specific timelines for outputs.

7

Importance of guidance on definition and prioritization of generative AI risks.

NIST should seek to support organizations by creating a taxonomy of risks specific to synthetic content — including this within their recent (draft) report on reducing risks posed by synthetic content (NIST AI 100-4) — or including specific sections on this within a broader risk and harm taxonomy. This could be supplemented by guidelines for risk prioritization, assessment, and decision-making, to help organizations better understand how to manage the possible risks and trade-offs related to synthetic content.

8

Documentation should be standardized and specific to different actors in the AI value chain.

Companies would benefit from more detailed guidance on what information should be shared to enable best management of risks across the AI value chain and ecosystem — or even standardized forms for reporting and sharing risk management activities, outcomes, and incidents, where practicable.

9 Companies are not clear on the best benchmarks and metrics to use for synthetic content detection, labeling or removal.

NIST could provide flexible guidance on detection and labeling of synthetic content, considering the evolving state of the art in synthetic content generation which includes a list of “recommended” metrics for measuring the success of mitigation activities.

10 Help needed to identify best tooling options for managing synthetic content risks.

NIST should invite participants from across the AI value chain to submit use cases specifically focused on tooling ecosystems, and provide guidance on best practices to support organizations in labeling and detecting synthetic content, verifying authenticity and tracking content origin.

Additional Considerations

During our analysis, another category of challenges emerged around resourcing of generative AI risk management efforts, specifically around workforce skilling issues and available budget. NIST could also offer guidance on optimizing and prioritizing risk management practices to account for any budget constraints and aligning efforts with specific areas where benefits compare favorably to costs.

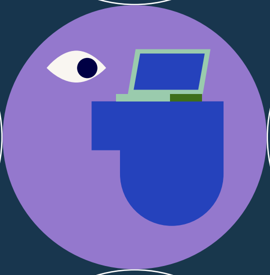
It may also be valuable to consider the challenges organizations may face in adopting the AI RMF, such as talent and skills shortages, and provide guidance on expected investments and training.

We included a discussion of these issues at the end of the main findings section as “additional considerations.” While perhaps out of scope for NIST, these issues are nevertheless barriers which may prevent companies from adopting or effectively implementing the NIST AI RMF and accompanying guidance on generative AI, and therefore should be acknowledged by the policy community and taken into account where possible.

Contents

Context

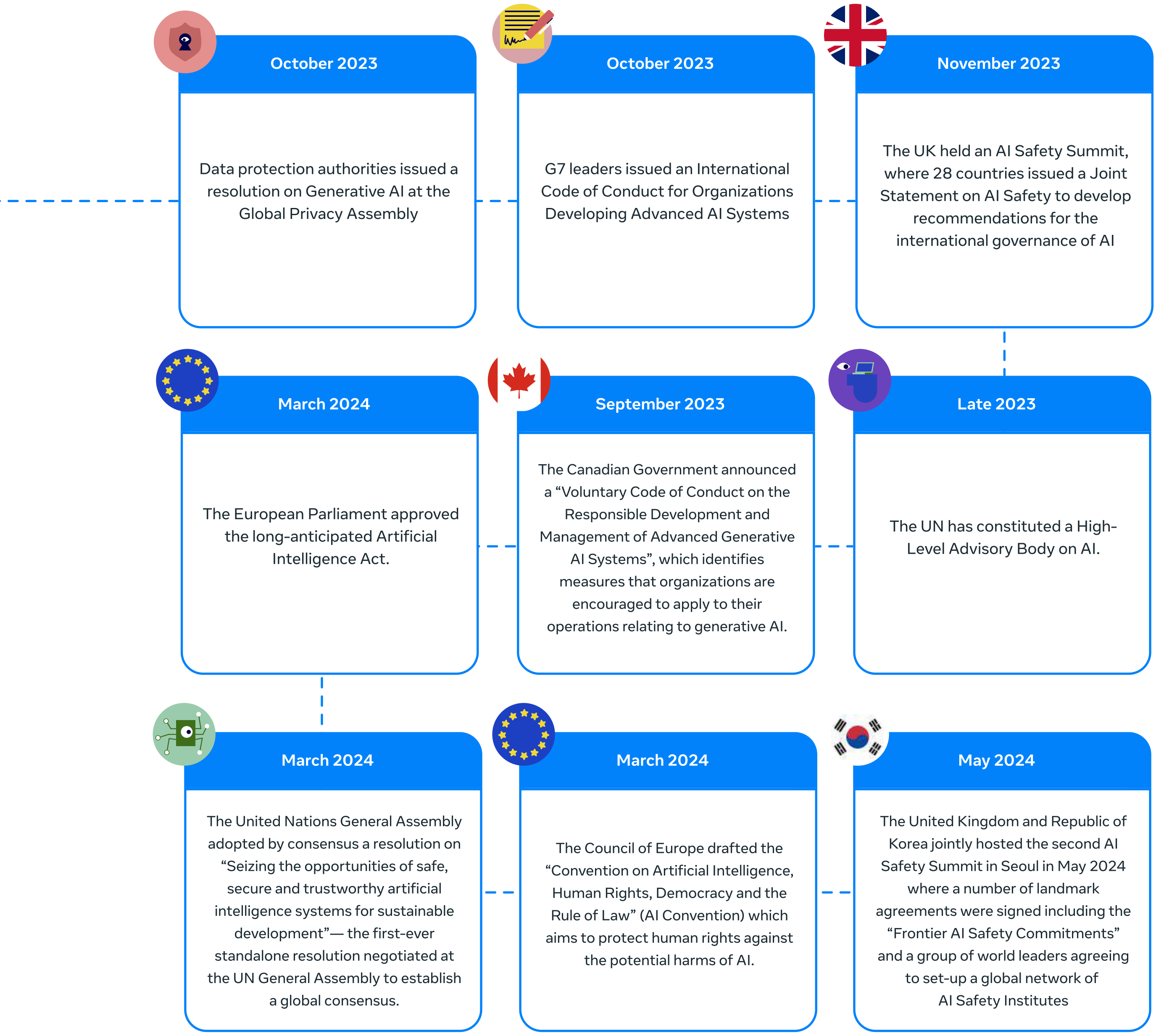
1



The policy context in the US and globally has been moving fast to keep pace with technological developments in generative AI and foundation models.

The US Executive Branch has demonstrated a significant investment in accelerating effective AI risk management, launching a set of voluntary commitments for AI in July 2023 for large model developers which created an initial set of principles for the safe development of generative AI foundation models. The White House issued a landmark Executive Order in October 2023 on “Safe, Secure, and Trustworthy Artificial Intelligence (AI).” The Executive Order sets a broad and ambitious agenda for the responsible development and use of AI in the US, with a focus on protecting citizens, promoting equitable outcomes, and ensuring the US remains a leader in AI innovation.

Beyond the US, we have also seen an intensification of activity among national, regional and multilateral institutions in producing both binding and non-binding guidance for generative AI:



October 2023

Data protection authorities issued a resolution on Generative AI at the Global Privacy Assembly

October 2023

G7 leaders issued an International Code of Conduct for Organizations Developing Advanced AI Systems

November 2023

The UK held an AI Safety Summit, where 28 countries issued a Joint Statement on AI Safety to develop recommendations for the international governance of AI

Late 2023

The UN has constituted a High-Level Advisory Body on AI.

September 2023

The Canadian Government announced a “Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems”, which identifies measures that organizations are encouraged to apply to their operations relating to generative AI.

March 2024

The European Parliament approved the long-anticipated Artificial Intelligence Act.

March 2024

The United Nations General Assembly adopted by consensus a resolution on “Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development”— the first-ever standalone resolution negotiated at the UN General Assembly to establish a global consensus.

March 2024

The Council of Europe drafted the “Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law” (AI Convention) which aims to protect human rights against the potential harms of AI.

May 2024

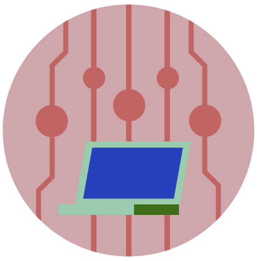
The United Kingdom and Republic of Korea jointly hosted the second AI Safety Summit in Seoul in May 2024 where a number of landmark agreements were signed including the “Frontier AI Safety Commitments” and a group of world leaders agreeing to set-up a global network of AI Safety Institutes

> These emerging guidelines and regulations are currently aligned around the need to focus on high-risk uses and apply flexible, industry-informed approaches to AI risk management.

These common threads are welcomed by industry, as divergences in approach or incompatible requirements between jurisdictions can impose high costs on enterprise and hamper innovation. Strategic and tactical differences between regulators can also thwart cooperation between international bodies, significantly undermining the capacity to identify and defend against cross-border risks and challenges that require mitigation and management across AI value chains, which are often global.

In this Open Loop program we endeavor to support the development of effective, evidence-based policy. We have gathered evidence on the current risk management practices of 40 companies who are developing and deploying generative AI to understand how the AI RMF 1.0 is supporting their efforts, and where further guidance is needed. With the insights in the following chapter we hope to support the development of national and international policy in this nascent and fast-evolving space and to help answer questions about how to best support companies on their risk management journey.

The definition of **red-teaming** used in this report

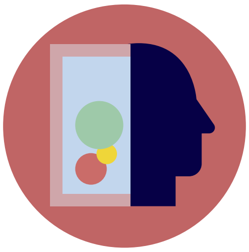


Red-teaming can be described as a practice — developed in the field of cybersecurity — of simulating a cyber-attack on an organization to test its defenses and identify vulnerabilities before they can be exploited by attackers. There is no single definition of, or approach to, red-teaming.

For the purposes of this report, we leverage the definition of AI red-teaming introduced by the White House Executive Order as: "A structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI. Artificial Intelligence red-teaming is most often performed by dedicated 'red teams' that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."[1]

Red-teaming for AI systems, and in particular for generative AI, is an evolving practice that encompasses a wide range of methods with varying levels of technical skills, access privileges and domain expertise required from the red-team. Past efforts have revealed that there is no one-size-fits-all approach, and highlighted the need for embedding red-teaming into a wider risk management effort. The process of red-teaming allows AI model developers and deployers to address flaws and vulnerabilities they find through the process.

The definition of **synthetic content** used in this report



Synthetic content generation is a broad term for the production of artificial or "synthetic" media artifacts by automated means using generative AI.

Other terms which are commonly used include "AI-generated content", "media generated by AI" and "synthetic media", and "AI-manipulated content" referring to the fact that AI-created images, audio and text are often built from "real" inputs which have been manipulated to achieve a particular effect. The wide accessibility of tools which facilitate the easy creation of synthetic content could lead to negative consequences, such as detriment to copyrighted brands and products, the manipulation of public opinion about a person, place or thing, erosion of trust in institutions or disruptions of democratic processes globally.[2]

While AI has been used to create "synthetic" media for several years, the use of generative AI tools has led to an increase in this type of media. Detection, authentication and labeling tools and techniques are being developed, alongside detailed frameworks and guidance such as Partnership on AI's Synthetic Media Framework.[3]

About the cohort

We had 40 participating companies in the program including AI startups, AI risk and assurance companies, and established multinational enterprises across various industries. Individual participants represented a diverse range of expertise, with both senior-level decision-makers and individuals involved in operational aspects of safety, compliance, and technology development. Each participant brought a unique perspective on AI risk management and the implementation of NIST's AI RMF. The multidisciplinary composition emphasizes the need for collaboration across various functions to address the multifaceted challenges of generative AI.

This chapter is split into two sections. In the **first (2.1)** we provide our findings relating to AI red-teaming, and in the **second section (2.2)** we set out our findings on synthetic content risk.

Findings & Policy Recommendations

Phase 1 of the Open Loop program employed a mixed methods approach to answering key questions surrounding implementation of generative AI red-teaming and synthetic content risk practices among companies **(Annex 1 for more details).** We focused particularly on the challenges faced by our cohort, and opportunities for addressing these. The findings that are presented in this section have been identified through online survey responses, case study submissions and two deep dive in-person workshops with participating organizations.

2

Overview

The organizations in our cohort expressed a clear motivation to implement AI risk management measures with the aim of building trust and meeting customer and regulator expectations. In our red-teaming workshop, participants emphasized their drive to facilitate "regulatory compliance in enforceable locations" and "manage brand reputation and customer experience." With these motivations in mind, companies were focused on practical barriers presented by the novelty of this risk management practice.

Where synthetic content was concerned, they similarly recognized the potential risks posed by misinformation, disinformation and toxic or inaccurate content being produced by AI and were eager to address these risks.

The sections below provide further detail on our findings, with each followed by an accompanying recommendation.

2.1 Findings and recommendations on red-teaming for generative AI

FINDING

2.1.1 AI risks and actors are not yet comprehensively mapped

DETAILS

As generative AI is still a nascent technology, the risks which might result from its use were not yet fully addressed by the group, and they reported feeling that they needed a more comprehensive mapping and categorization of potential risks and the conditions under which they could arise. Understanding types of risk and their potential vectors was seen as crucial to designing appropriate systems for managing risks.

In addition, our expert group noted that the scope of risks to be considered is broad, and that it isn't consistently made clear that when talking about risks, this should include customer and user risks, not just company risks.

International efforts to create aligned taxonomies and terminologies for traditional AI in which NIST has participated could provide a good basis for this type of intervention. The process and working practices established in particular for the EU-US Trade and Technology Council (TTC) Working Group 1 (WG1) sub-group on AI Taxonomy and Terminology could prove informative.

Other instructive examples may be NIST's own *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*[5], the "Taxonomy of Trustworthiness for Artificial Intelligence" and the detailed risk profiles developed for general purpose AI systems (GPAIS) by experts at UC Berkeley's Center for Long-Term Cybersecurity (CLTC)[6].

Note: NIST have embarked upon the development of this taxonomy within "NIST AI 600-1 (draft)"[7] and have taken an important first step, however the taxonomy requires further elaboration in terms of providing technical definitions of terms used. International collaboration to create alignment could further enhance the impact of this work.

Somewhat relatedly in their efforts to understand and articulate the foundational aspects of the generative AI ecosystem and value chain, the companies expressed a need for a definition of the roles and responsibilities of different actors across the AI development lifecycle and value chain. An AI system needs to be tested at multiple stages across development, deployment, and maintenance. For example, the question of whether a generative AI foundation model that is tested by the developer, must be tested again by downstream organizations deploying it as part of a customer service chatbot.

To emphasize these interconnections, one workshop participant suggested that, without clearly defined roles, the burden of providing the resources for red-teaming may fall disproportionately at one point or another along the value chain.

We heard of similar challenges from companies specifically relating to defining risks, roles and responsibilities where synthetic content generation is the use case, and feel that the below recommendation would help to resolve these challenges for both topics **[see 2.2.1].**

2.1.1 Enterprise-level metrics enable organizations to consistently assess quality and success

RECOMMENDATION

2.1.1

There was strong support for the development of a taxonomy of risks within our cohort, with participants at the workshop rating it a top priority for development by NIST (now partially addressed by NIST AI 600-1) . Furthermore, NIST should identify and define the roles of different actors in the AI value chain to support organizations in their strategic and tactical efforts to establish and/or scale generative AI risk management systems, ensuring that they focus their efforts on the areas and tasks which they must prioritized given their role in the value chain. Practical guidance for the application of risk taxonomies will also be essential, as not all risks will be applicable to all systems. Guidance to support organizations system categorization (like FIPS 199) can help organizations target their more detailed risk assessments.

Note: In particular we hope that our suggestion on systems categorization helps to answer NISTs questions on whether further categorization of risks and systems is needed, and whether further detail should and could be provided to Section 3 of NIST AI 600-1 (draft). In general, while the table of risks and attendant mitigations is very helpful, companies may still struggle to define what "good" looks like in the establishment and maintenance of such systems and policies without more detailed descriptions. For example, in NIST AI 600-1 (draft) at GV 1.1 — 002: "Define and communicate organizational access to GAI through management, legal, and compliance functions." Such instruction is useful, however companies would further benefit from accompanying guidance on what such a communication protocol should look like, how it should be maintained, contingency planning in case of staff turnover etc.

FINDING

2.1.2 Red-teaming guidelines could provide enhanced practical support

CHALLENGES

Our cohort raised a number of implementation challenges that new generative AI red-teaming guidance could help to address:

Purpose and scope of AI red-teaming exercises

There are many aspects of AI systems which can be tested via red-teaming efforts, to measure and help improve security, reliability and safety. Defining an appropriate red-teaming exercise for a system depends upon multiple factors such as the context of the system and its foreseen uses. Where the companies are facing constraints around resources, they would value guidance on when red-teaming might be the most appropriate mitigation, and when it requires supplementation with other techniques.

To ensure effectiveness, clear goals are required including defining what kind of adversary the attack will simulate, which types of attacks the red-team will simulate, and which vulnerabilities the exercise(s) will seek to uncover.

Given the breadth of possible goals, companies would value guidance from NIST on which risks AI red-teaming is best suited to managing in the context of generative AI, as well as how red-teaming should be integrated within the overall risk management strategy. The guidance should provide support in defining the purpose and scope of red-teaming exercises, encompassing structured techniques and the relationship to other generative AI risk management activities.

Composition of AI red teams

An area which received particular attention among our participants was around the composition and skills needed for an effective AI red-team.

Clarity on roles, responsibilities, and necessary training for red team members to maintain objectivity and integrity in the red-teaming process was seen as important. Participants highlighted that it would be useful to have high-level guidance on the structure, number of individuals, and blend of skill sets, required to complete the red-teaming activities appropriate to their context and use cases. The specialized skill set required for effective AI red-teaming presents a significant challenge, with many companies lacking the internal expertise to build and manage in-house red-teaming teams.

The need for a multidisciplinary approach was however expressed, with one participant, who represented a company that performed red-teaming for others, noting that; “What we often see our customers asking for is expert knowledge regarding the AI product and what it does, but also having the adversarial mindset that is required for more sophisticated attack vectors and being able to explore and access them. The goal is to combine these skill sets - often this is across multiple people.” The cohort also discussed bringing technical skills together with insights from social sciences, risk management and domain experts, to bridge the gap between technical testing and understanding the material impact on people and society. Selected case studies highlighting the variation in approaches to red-teaming composition are presented in Annex 3 (ref. 1a).

External, internal and automated AI red-teaming

Companies expressed challenges in integrating external red-teaming and a desire for guidance to help address coordination challenges and ensure alignment with internal processes. This could include facilitating information-sharing mechanisms to enhance collaboration with external red teams and defining protocols for transparent communication and integration of external red-teaming outcomes into the AI development lifecycle.

Choosing between internal and external red-teams presents another key tension for effective generative AI red-teaming. Internal teams may lack the necessary expertise and fresh perspectives often brought by external red-teams. Furthermore, building internal expertise requires investment in training and may take considerable time. On the other hand, participants noted that external red-teams can be expensive, and their lack of familiarity with the specific AI system or use case and company context may require additional onboarding time. Further, in some cases, external red teams may be less experienced with adversarial techniques and approaches compared to more experienced internal red teams. When asked whether they were performing red-teaming in-house, outsourcing the program, or combining the two, the majority of participants indicated that their approach would be a combination.

Companies were also aware of the importance of automation in scaling red-teaming efforts and sought guidance on making them more measurable and reliable, and limit human exposure to toxic content. NIST should encourage the use of automation technologies to scale AI safety evaluation and risk mitigation by providing guidelines on how automated red-teaming can complement manual red-teaming efforts.

☆ RECOMMENDATION

2.1.2

In developing guidance, NIST should seek to support organizations in establishing the purpose and scope of generative AI red-teaming efforts, by providing a systems engineering framework for managing risk in AI systems (akin to NIST 800-160r1). That framework should articulate the appropriate use of detailed AI red teaming and scaled/automated measurement, and describe the basic steps that make up an assessment (similar to what is presented in NIST 800-30r1).

NIST should — developing further upon the work done in the draft “Gen AI profile” — define guidelines on red team composition, emphasizing the importance of independence from the safety or development team. Ideally, this should include recommendations for ensuring diversity in red team composition, including technical experts, domain specialists, and individuals with diverse backgrounds and perspectives, with consideration to organizations’ varied sizes and resources. It should also provide guidance on what staffing is necessary to maintain in-house as opposed to external support, and break down the roles necessary for system assessment, management, authorization, etc. (akin to “Roles and Responsibilities” described in NIST 800-137). This would ensure that their guidelines reflect the comprehensive suite of challenges organizations currently face in embedding robust red-teaming practices.

FINDING

2.1.3 Enterprise-level metrics enable organizations to consistently assess quality and success

DETAILS

Building on the requirement for a clear scope for red-teaming activities, a key topic that was explored in the workshop was the value of NIST developing a detailed suite of metrics for assessing the effectiveness of red-teaming for generative AI systems. Many participants said a lack of defined metrics and tools was a significant barrier to measuring the effectiveness of red-teaming, discussing issues such as what aspects of the AI system to test, when to conduct testing throughout the development cycle, and how to structure the testing approach. Participants spoke about the degree of expertise required to establish the right metrics for each use case, describing "unknown unknowns", as challenges to successful red-teaming exercises. In addition, our expert group raised that there might be different levels of red-teaming needed based on the product and its impact, and that testing should be context specific.

However, in considering the range of possible metrics that could be considered at the system level, and the requirement to tailor them to specific applications or use cases, participants highlighted the challenge of setting success metrics with an appropriate level of granularity. Metrics set at the specific risk, use case, or red-teaming activity level may not have the broad applicability required to support cross-industry best practices.

The difficulty of setting a single suite of metrics was emphasized by the range of use cases submitted by our cohort. Whilst there are general risks related to generative AI such as the inaccuracy of outputs and the vulnerability of sensitive data, there is a requirement to connect metrics of this type to the possible real-world impacts. Supporting organizations in assessing the identified risks requires metrics which go beyond simply counting vulnerabilities and consider the impact on fairness, safety, and privacy.

With regard to securing executive buy-in and building organizational awareness about AI risks, participants highlighted the requirement for readily available guidance on how to evaluate the success and impact of a red-teaming program. This included which metrics to set at a program level, such as the number of vulnerabilities identified, the effectiveness of mitigation strategies, and the overall improvement in the safety and robustness of generative AI systems.

These enterprise-level metrics are at a higher level of abstraction, and therefore could be more standardized, than use case level reporting. To understand how our cohort were currently thinking about assessment criteria for red-teaming efforts, we asked participants to consider how they would define a successful program. Two themes emerged, encompassing business- and people-centric outcomes.

Business-centric outcomes

Including meeting ethical commitments, time-efficient red-teaming and remediation activities, mitigating reputation risk, minimizing risk of regulatory penalties, minimizing additional costs through automation, developing testing programs that can be rerun at regular intervals for scaled and ongoing verification, avoiding detrimental impacts on product release cycles, and increasing the rate at which clients launch AI-powered services.

People-centric outcomes

Including building trust with consumers and stakeholders, increased confidence amongst users and developers, improved team security culture, creating awareness of red-teaming as an option, educating teams on realistic adverse scenarios, and developing the ability to share red-teaming efforts with customers.

2.1.3 Enterprise-level metrics enable organizations to consistently assess quality and success

☆ RECOMMENDATION 2.1.3

NIST could account for the need to demonstrate the value and impact of red-teaming programs by producing indicative enterprise-level metrics and quality standards. As part of this, NIST should consider driving and supporting the setting of benchmarks by international standards organizations by gathering data on what the benchmarks should be, and comparing and evaluating emerging benchmarks. As benchmarks are established and widely adopted, NIST may also be able to provide authoritative training data sets that can help organizations fine tune their systems or otherwise align with commonly accepted measurements.

In NIST AI 100-5 ("A Plan for Global Engagement on AI Standards") the organization emphasizes the need for standardized protocols for red-teaming; we hope that this recommendation provides some additional direction on how this could be developed.

Additionally, our expert group recommended that NIST should support efforts to communicate with and educate the public on red-teaming, including topics such as the difference between public harm and products not working as intended. This work could help to bridge the gap between business value and societal value by connecting business risk outcomes to real-world impacts.

FINDING

2.1.4 Case studies and best practices on AI red-teaming are valuable

DETAILS

In discussing what support organizations need to embed scaled and effective generative AI red-teaming practices, our cohort highlighted the importance of openly available case studies showcasing effective implementation across market demographics. Participants expressed a desire for a place to share "best practices, experience, or misuse scenarios" and "anonymous reporting and tracking of risks…and vulnerabilities." **Annex 3 (ref. 1b)** provides an example of an end-to-end red-teaming process implemented by a contributor representing an AI governance company.

☆ RECOMMENDATION 2.1.4

To understand current implementation practices, and develop a view of best practice, NIST could solicit detailed case studies which give insight into how organizations are setting their approach to red-teaming in practice. It may be valuable to establish a central online platform for sharing resources (e.g., case studies, best practices guides) and fostering collaboration among stakeholders in the AI red-teaming community. This could be modeled after the National Cybersecurity Center of Excellence (NCCoE), which provides a space for organizations with common challenges to articulate requirements for solutions and illustrate opportunities for innovation.

FINDING

2.1.5 There is demand for open source tools and techniques that enable efficient AI red-teaming

DETAILS

In assessing the tools available for performing generative AI red-teaming, the strongest theme that emerged from the workshops was the opacity of the market. It is not currently very clear what is available (particularly open-source tools), what capabilities these tools have, and which are required for a given application. A lack of confidence in the available tools was a consistent theme, with another participant noting that "even if you want to implement basic transparency tools, there is often a struggle with compatibility with existing systems. Sometimes we're not sure if tools are really trusted tools or something else".

A number of workshop participants described using a blended tooling approach, with one noting that "we've found a lot of excel spreadsheets throughout the industry. We've tried using automated tooling. One client is adding a chrome extension to log what red teams are doing. There are a lot of homegrown solutions. One tool used for logging interactions would be useful." This reflects one of the main issues raised by our expert group, who noted that a significant challenge of generative AI is that it is impossible to exhaustively identify all failure points of a giant model, and therefore red-teaming should not be thought of as a single exercise.

The role of NIST in driving the adoption of specific tools for generative AI red-teaming was a key discussion point. Some participants questioned the feasibility of this task for a single institution, considering the diverse organizational requirements. Others expressed concern that NIST's influence might unintentionally limit tooling options. This could prematurely impact the market for tools in this early phase of development, and the participating organization noted that, "we don't want monopoly control, and small companies might not have resources to use some tools".

Participants evaluated encouraging the development of open source tools as a future option for the field, with the advantages of reducing the likelihood of a monopoly and enabling smaller organizations to access best practice tools.

☆ RECOMMENDATION 2.1.5

NIST should encourage transparency regarding available tools, conducting a stocktaking exercise and classifying those which are currently available by capability and availability, with a view of which red-teaming activities a given tool would be appropriate for. This could support organizations in selecting tools, without exerting undue influence over tool development. Guidance already suggested in this document for benchmarking, evaluation, and formalized risk assessment processes would also encourage further standardization of acceptable practice in the ecosystem; likely facilitating more comparability between open tools and a more competitive, informed marketplace for their use.

FINDING

2.1.6 Frameworks and guidelines which do not interoperate are a barrier

DETAILS

Our cohort felt that guidelines should be interoperable and flexible given the fast pace of change in the AI space. There has been increased regulatory emphasis on the role of red-teaming in AI risk management. For example, the EU AI Act requires providers of general purpose AI models with systemic risk to conduct and document adversarial testing of models[8], the UK government identifies red-teaming as an emerging process for frontier AI safety[9], and the Hiroshima Process International Guiding Principles and Code of Conduct highlight the need for AI developers and deployers to include red-teaming exercises in their risk management programs.[10]

While red-teaming is featured in the NIST AI RMF Playbook[11], it is only mentioned as a means of testing a system's security and resilience in a traditional cybersecurity context. NIST has already however gone some way to addressing this within their (draft) "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile" where red-teaming is described and mapped to potential risks which it can be effective in mitigating.

To add to this, we learned from the program companies that effective guidance would need to take in the requirements and recommendations from the broader landscape and build on these to facilitate an interoperable approach, encouraging organizations to pursue a wide panel of goals, such as testing the model's robustness and reliability, and its potential to issue toxic or biased content.[12]

☆ RECOMMENDATION

2.1.6

NIST should continue to leverage existing guidance and definitions provided by multilateral organizations (e.g. the OECD) to facilitate harmonization of guidance across jurisdictions as they have begun with their work on crosswalks, and develop a roadmap for global engagement to promote AI technical standards, ensuring that NIST's guidance is informed by relevant technical standards and investing in mechanisms for coordination with international peers.

We welcome the recently published (draft) "[A] Plan for Global Engagement on AI Standards" and encourage NIST to provide more detail on their anticipated timeline for the development of international consensus standards through the outlined "high priority implementation actions specific to the U.S. government". While this may be challenging as there are many factors to consider, many of which are outside of NISTs control, even an indicative timeline (specifying the minimum time they expect will be required) will be useful to industry in terms of informing product development roadmaps and investment decisions relating to long-term projects and investments.

FINDING

Additional consideration: **The cost of AI red-teaming programs and resource limitations**

DETAILS

The fast-paced, resource-constrained environment of startups and small and medium-sized enterprises (SMEs) may present unique challenges in implementing structured AI red-teaming practices. Startups and SMEs may lack the dedicated personnel and budget resources to conduct comprehensive red-teaming exercises. One workshop participant noted "There are never enough resources for things like red-teaming and other security efforts. Cost is also important - if you tell a CEO we need to form a red team that will cost USD "X" and will last forever, that will get shut down. We need to find a way to be budget neutral." Other participants also raised the question of cost, highlighting that it could be valuable to have greater transparency around the scale of costs for different types of red-teaming activities.

Participants frequently cited budget limitations as a major hurdle. Insufficient funds restrict a company's ability to invest in several crucial aspects of AI red-teaming. These limitations include hiring skilled red-teaming professionals with expertise in AI security and adversarial techniques, and the acquisition of necessary red-teaming tools, platforms, and infrastructure, potentially hindering the efficiency and effectiveness of testing efforts. Limited budgets may also restrict the scope and frequency of red-teaming exercises, leading to potentially incomplete risk assessments or unacceptable time gaps in such assessments. A participant in the workshops suggested that, to address this, there might be a special program for startups under a certain business value or which meet certain conditions such as not having an angel investor.

OPPORTUNITY

A.C

NIST cannot impact the budget or investment capacity of an organization, but in order to achieve widespread adoption of effective red-teaming, NIST could encourage collaborative models, resource-sharing mechanisms, and the use of open source tools to facilitate consistent red-teaming implementation and help address skill and budget constraints. Participants and our expert group highlighted that it may be necessary to produce specific recommendations for different groups such as those starting from scratch versus those with established practices, and startups / SMEs vs large organizations, to address their unique challenges and resource constraints. Profiles of risk assessment frameworks for SMBs, like NIST has created for the Cybersecurity RMF and the CSF, will be very helpful in supporting the breadth of organizations we expect to adopt GenAI tools.

2.2 Findings and recommendations on Synthetic Content Risk

Here we present our main findings on synthetic content risk management. We have followed the same structure as above with each of the main findings followed by a recommendation for NIST.

FINDING

2.2.1 **Importance of guidance on definition and prioritization of synthetic content risks.**

DETAILS

As with red-teaming, the companies were seeking support in defining synthetic content risks, and the roles and responsibilities of different actors in the value chain **[see 2.1.1].**

Throughout the workshops, participants highlighted the lack of complete or clear guidance regarding which risks to target and how to define an acceptable level of risk as one of the challenges to developing effective policies, risk assessments and procedures. Beyond the requirement for a taxonomy of risks, there was also a call for guidance on risk prioritization, particularly where trade-offs are required in designing mitigation strategies. NIST has substantively met the need for this guidance in their recently published draft guidance: “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile”. However, below we articulate within our recommendations how NIST may build on this strong foundation to even more comprehensively meet this need.

Uniquely in the conversations on synthetic content risk there was the concern around possible trade-offs required when working towards transparency or direct disclosure of synthetic content as a goal. For example, one of the companies in the cohort spoke of balancing transparency and data protection — they were concerned that in their efforts to be fully transparent about the provenance of a piece of synthetic content that they may inadvertently reveal data about the publisher, such as location or other personal or sensitive data. Similarly, there is tension between making generative AI models more publicly available so that groups are able to assess the outputs, and limiting access to deter model misuse.

Recognizing that not all synthetic content poses the same risks, participants noted that their companies were already conducting risk assessments to prioritize resources and tailor mitigation strategies. Annex 3 (ref. 2a) introduces a case study on generative AI in the aviation industry which gives an insight into the spectrum of risks under consideration, including safety, security, operational effectiveness, legal compliance, and the industry's reputation. The case study also highlights the role of content labeling and authentication in supporting informed decision-making and reducing the risk of misinterpretation or misuse.

2.2.1 Importance of guidance on definition and prioritization of synthetic content risks.

RECOMMENDATION

NIST should seek to support organizations by creating a taxonomy of risks specific to synthetic content, or including specific sections on this within a broader risk and harm taxonomy. This could be supplemented by guidelines for risk prioritization, assessment, and decision-making, to help organizations better understand how to manage the possible risks and trade-offs related to synthetic content.

Also, by adopting a risk-based approach and focusing on scenarios where confusion about content origin poses significant harm, organizations can better address the challenges of synthetic content. By distinguishing between content that is realistic enough to mislead users and content that is obviously fictional, NIST can tailor risk mitigation strategies to address specific challenges effectively.

Note: We acknowledge the work that NIST has done to provide this risk taxonomy within the "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile", and accompanying mitigations. However, no such corresponding list has been provided within "Reducing Risks Posed by Synthetic Content An Overview of Technical Approaches to Digital Content Transparency", and we recommend that NIST directly reproduce, using the same language, those risks identified within the list of twelve given in the "Gen AI Profile'' paper (i.e. Human-AI Configuration, Information Integrity) in the Synthetic Content guidance, so that it is clear and explicit to companies how these pieces of guidance intersect and reinforce one another. This should also support companies prioritize foundational tasks for AI risk management.

FINDING

2.2.2 Documentation should be specific to different actors

DETAILS

Another challenge participants indicated was that of the responsibilities of each company within the AI value chain, and collaboration between AI actors towards achieving transparency and disclosure (labeling) goals.

The focus of the discussion was on the need for clearly defined documentation standards across the chain, with particular emphasis on strong collaboration around transparency goals between upstream and downstream AI actors. Our expert group advised NIST to create more clarity around roles and responsibilities, considering accessibility of various disclosure methods, accountability mechanisms and responsibilities for stakeholders along the AI value chain.

Participants noted that "having clear expectations for others in the AI chain would help us to determine what level of trust we could put into third party services," and "addition of service-level agreement and data contracts can be valuable to provide technical definition on the implementation of these transparent data exchange systems which can be fed by synthetic content." While some participants noted that open source models with available model weights allows AI actors to better empower downstream developers to take control on the transparency efforts, others declared these were a "very low priority in a startup - we trust that vendors of generative AI products will have taken care of this upstream".

☆ RECOMMENDATION

NIST could develop a suite of template risk assessments and specific documentation types for different actors in the AI value chain. Also setting expectations on information sharing between actors across the AI value chain would foster transparent collaboration, as noted by participants to our workshop.

Note: We welcome the inclusion of risk documentation and techniques by NIST in the (currently draft) *"Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile"*. We would encourage NIST to build upon this to further specify appropriate or standardized AI actor-specific documentation formats, so that incidents can be analyzed at an appropriate level depending upon the role of the actor in the value chain, and future incidents better mitigated against.

FINDING

2.2.3 Companies are not clear on the best benchmarks and metrics to use

DETAILS

One of the challenges raised was that of measuring effectiveness of implemented methods, with regards to the transparency goals and requirements, with participants noting, for example, that "Benchmarks such as clear KPI's would be useful." In particular, participants highlighted the need for clear benchmarks to evaluate both robustness of methods and post-implementation effectiveness. Participants also noted that guidance on how to use the metrics would be helpful, "as well as who should be seeing the metrics," highlighting the challenge that even with the appropriate metrics identified, interpreting them and using them to guide decision-making processes also requires new guidance.

With regard to metrics, NIST could explore establishing thresholds and risk levels for various content categories to support consistent approaches across organizations. However, participants again highlighted the need for contextualizing guidance to industries, sectors, geographies and organization sizes, noting that "metrics can be useful if they are also considering geography, size and industry segment factors," and "useful but also usually at industry level". Furthermore, NIST could consider accompanying the guidance on metrics and thresholds with supporting documentation on how to use these and who should be involved in the review process.

☆ RECOMMENDATION

2.2.3

NIST could provide flexible guidance on detection and labeling of synthetic content, considering the evolving state of the art in synthetic content generation which includes a list of "recommended" metrics for measuring the success of mitigation activities.

NIST should also leverage the existing framework and guidelines produced by "Partnership on AI" (PAI) in their Synthetic Media Framework. This could help organizations along the AI value chain mitigate risks posed by synthetic content by considering the applications and limitations of technical approaches for disclosure, such as watermarking, cryptographic metadata, and fingerprinting, and promoting further research in this area. In addition, NIST could clarify the utility of using direct or indirect disclosure methods at the foundation model level or at the fine-tuned application level.

FINDING

2.2.4 **Help needed to identify best tooling options for managing synthetic content risks**

DETAILS

Participants in our cohort discussed a range of possible tooling for labeling, detecting and authenticating synthetic content, and establishing content provenance, however it is clear that there is limited consensus on which technologies, tools or tool stacks constitute best practice or offer the most reliable results. **Annex 3 (ref. 2b)** introduces five use cases submitted by our contributors to highlight the range of tools and combinations of tools being used to give a degree of transparency. The scope of technologies considered included watermarking, metadata tagging, blockchain-based solutions, visual indicators and textual disclaimers.

☆ RECOMMENDATION 2.2.4

NIST should invite companies from across the AI ecosystem to submit use cases specifically focused on risk management tooling, and provide guidance on best practices to support organizations in labeling and detecting synthetic content, verifying authenticity and tracking content origin.

Note: We welcome the provision of such a list in the (currently draft) guidance from NIST on *"Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency"* and would encourage NIST to even further elaborate upon this list by drafting companion playbook guidance which indicates to small companies with limited resources how they may conduct a risk assessment for synthetic content and mitigate for these risks in a step-by-step manner with illustrative examples.

FINDING

Additional consideration: **Provide free training materials or provide guidance around expected investments in training on synthetic content risk management**

DETAILS

The field of synthetic content risk is continuously evolving, with new techniques being developed at a fast pace, and companies sometimes struggle to keep their workforce and AI practices up to date.

Most of the companies on our program were only in the initial stages of understanding the various methods and techniques available for the detection and authentication of synthetic digital content. With methods such as labeling, watermarking and fingerprinting only now coming to the fore, organizations are looking to upskill towards the implementation of risk management measures within their products. One participant at the workshops expressed concern around the ability of his colleagues to learn and implement these new techniques — "[A] key topic in terms of what kinds of skills are most needed - do these need to be internalized to organizations?"

The location in the value chain, or organization size can also be elements linked to limited technical capabilities or resources on these specific topics.

Participants on the Open Loop program saw the opportunity to support upskilling of the workforce through free online courses — "promote free coursework on the topics so large teams and even small organizations have an equal playing field of learning the skills".

OPPORTUNITY A.C

NIST can help organizations address cost as a barrier by providing free training materials on synthetic content risk methods. Participants to the workshops highlighted that "NIST support on putting this at the forefront is key."

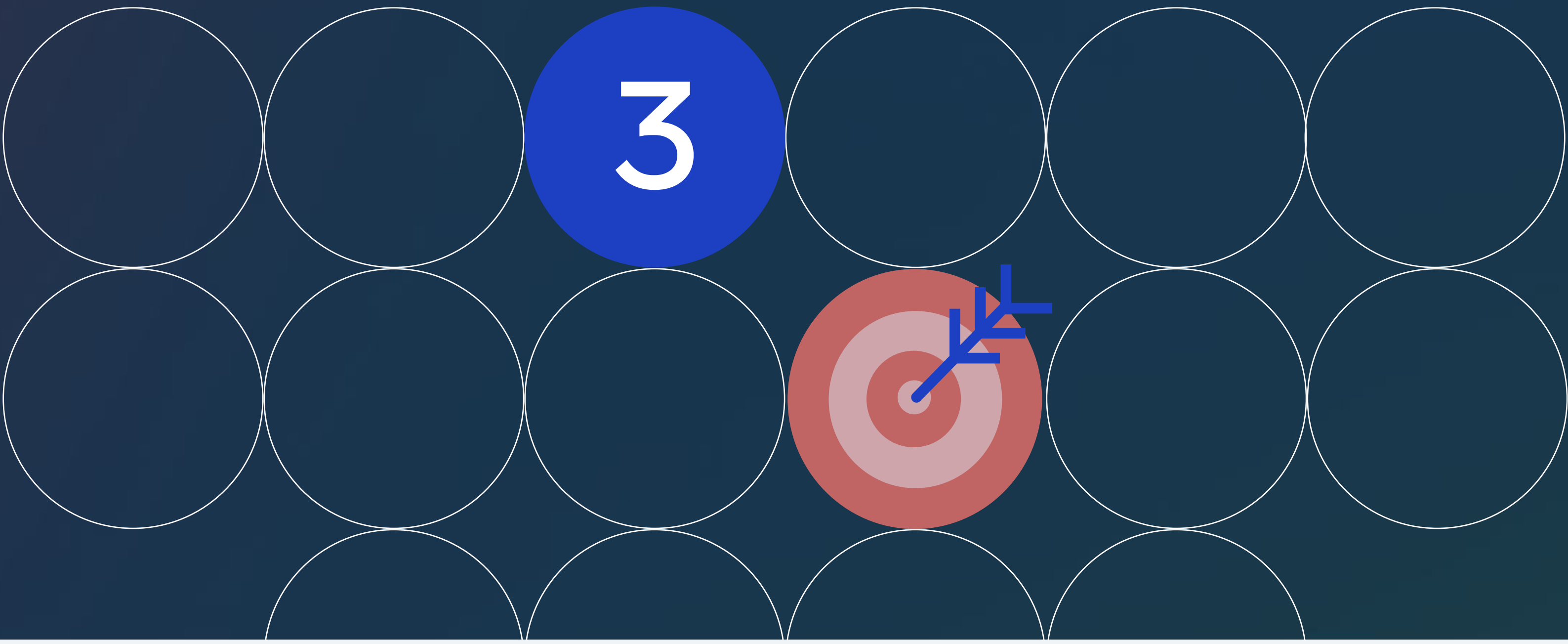
In this chapter we have outlined the findings from Phase 1 of our research and have also included some corresponding recommendations.

Conclusion and Next Steps: Phase 2 of the Open Loop US Program

This report marks the culmination of the first phase of the Open Loop US Generative AI Risk Management Policy Prototyping Program, providing a comprehensive stocktaking and analysis of emerging practices related to AI red-teaming and synthetic content risk management in the context of generative AI. Our findings underscore the importance of clear, practical, and future-proof guidance from NIST, highlighting the need for a detailed taxonomy of risks, tooling, and success metrics. The recommendations outlined in this report aim to address these needs, providing NIST with a roadmap to enhance the clarity, effectiveness, and accessibility of their upcoming guidelines on these areas.

As we transition into the second phase of the program, we will build upon the insights gained from this initial phase. The focus will shift towards understanding how companies might adopt and use the NIST AI RMF for generative AI risk management, and where there are opportunities for expanding the current provisions within the framework. In this work, we will take into account the activities already undertaken by NIST within their “Generative Artificial Intelligence Profile” (NIST AI 600-1) to provide supplementary and generative AI specific measures to the framework subcategories.[13]

We will delve into company practices and journeys, evaluate the AI RMF against policy prototyping criteria, identify challenges, generate recommendations for future AI RMF iterations, and examine the awareness and usability of NIST resources. This iterative approach will ensure that the lessons learned from Phase 1 inform the testing of the NIST AI RMF in Phase 2, providing a robust foundation for policy development and enhancing the effectiveness of the NIST AI RMF. We look forward to continuing this important work and contributing to the development of effective and evidence-based policies for generative AI risk management.



Bibliography

Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models. UC Berkeley Center for Long Term Cybersecurity, 2023. https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/.

Center for Data Innovation. (2024). https://www2.datainnovation.org/2024-nist-ai-eo-rfi.pdf
ITI. (2024). ITI Feedback to Request for Information (RFI) Related to the National Institute of Standards and Technology's (NIST) Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11) https://www.itic.org/documents/artificial-intelligence/ITIFeedbackNISTAIEORFIFINAL.pdf

EU AI Act, Text voted on March 13, 2024 by the European Parliament. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

European Parliament. (2023a). Generative AI and Watermarking. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf

G7. (2023). Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. https://www.mofa.go.jp/files/100573473.pdf

Heikkilä, M. (2022). How AI-generated text is poisoning the internet. MIT Technology Review. https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/GPAI, The Global Partnership on Artificial Intelligence. (2023). State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition for Public Release. https://gpai.ai/projects/responsible-ai/social-media-governance/Social%20Media%20Governance%20Project%20-%20July%202023.pdf

National Institute of Standards and Technology (NIST). (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

National Institute for Standards and Technology (NIST). (2023). NIST AI RMF Playbook. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

National Institute of Standards and Technology (NIST). (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. [page 11.] https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

National Institute of Standards and Technology (NIST). (2024). Adversarial Machine Learning. A Taxonomy and Terminology of Attacks and Mitigations. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

Responsible Practices for Synthetic Media Framework by Partnership on AI. https://syntheticmedia.partnershiponai.org/

The White House. (2023). Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. 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/

UK Government, Department for Science, Innovation and Technology (2023). Emerging processes for frontier AI safety. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety

References

[1] The White House. (2023). Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. 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] European Parliament. (2023a). Generative AI and Watermarking. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf
Heikkilä, M. (2022). How AI-generated text is poisoning the internet. MIT Technology Review. https://www.technologyreview.com/2022/12/20/1065667/how-ai-generated-text-is-poisoning-the-internet/GPAI, The Global Partnership on Artificial Intelligence. (2023). State-of-the-art Foundation AI Models Should be Accompanied by Detection Mechanisms as a Condition for Public Release. https://gpai.ai/projects/responsible-ai/social-media-governance/Social%20Media%20Governance%20Project%20-%20July%202023.pdf

[3] Responsible Practices for Synthetic Media Framework by Partnership on AI. https://syntheticmedia.partnershiponai.org/

4 Some AI risk management challenges facing companies will likely fall outside of NIST's remit. Nevertheless, we have included discussion of these where we see an opportunity for NIST to take any action which might improve the operating environment, for example, through interagency cooperation or support for industry efforts.

[5] National Institute of Standards and Technology (NIST). (2024). Adversarial Machine Learning. A Taxonomy and Terminology of Attacks and Mitigations. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

[6] Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models. UC Berkeley Center for Long Term Cybersecurity, 2023. https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/.

[7] National Institute of Standards and Technology (NIST). (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

[8] EU AI Act, Text voted on March 13, 2024 by the European Parliament. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

[9] UK Government, Department for Science, Innovation and Technology (2023). Emerging processes for frontier AI safety. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety

[10] G7. (2023). Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. https://www.mofa.go.jp/files/100573473.pdf

[11] National Institute for Standards and Technology (NIST). (2023). NIST AI RMF Playbook. https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

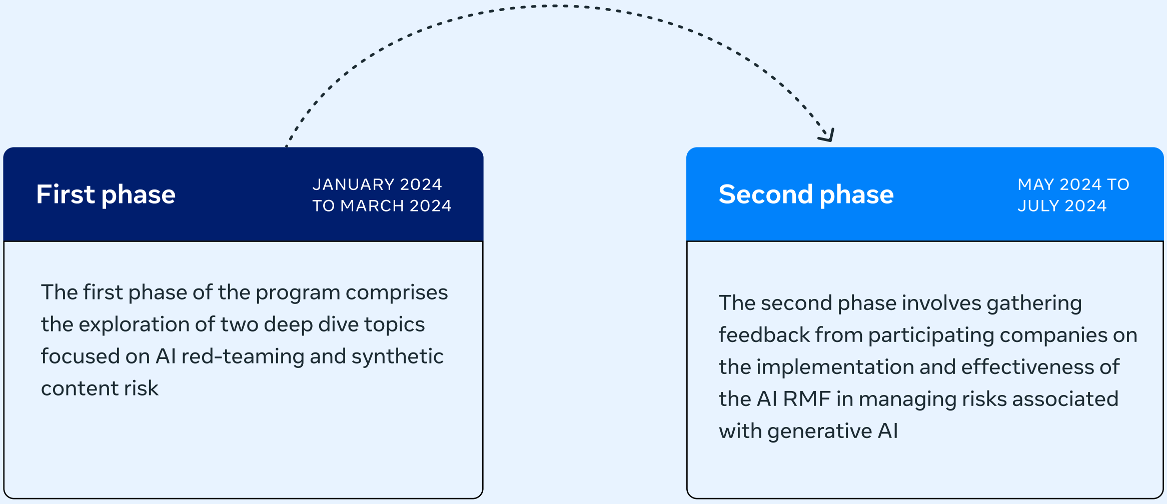
[12] Center for Data Innovation. (2024). https://www2.datainnovation.org/2024-nist-ai-eo-rfi.pdf
ITI. (2024). ITI Feedback to Request for Information (RFI) Related to the National Institute of Standards and Technology's (NIST) Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11) https://www.itic.org/documents/artificial-intelligence/ITIFeedbackNISTAIEORFIFINAL.pdf

[13] National Institute of Standards and Technology (NIST). (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. [page 11.] https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

Red-Teaming & Synthetic Content Annex

Annex - Methodology for Phase 1

The Open Loop US program comprises two phases:



First phase JANUARY 2024 TO MARCH 2024

The first phase of the program comprises the exploration of two deep dive topics focused on AI red-teaming and synthetic content risk

Second phase MAY 2024 TO JULY 2024

The second phase involves gathering feedback from participating companies on the implementation and effectiveness of the AI RMF in managing risks associated with generative AI

Phase 1

The results of phase 1 offer valuable insights into the current state of red-teaming and synthetic content risk for generative AI, highlighting challenges and opportunities for improvement. While acknowledging limitations, these findings lay the groundwork for future research, policy development, and the continued advancement of effective risk management practices in the rapidly evolving domain of generative AI.

Phase 1 was guided by the following key overarching research questions:

- **Question 1:** What are the current practices and approaches organizations are using for red-teaming and synthetic content risk management for generative AI?
- **Question 2:** What are the key challenges and best practices organizations are encountering in implementing red-teaming and synthetic content risk management for generative AI?
- **Question 3:** How are organizations utilizing the NIST AI RMF to guide their red-teaming and synthetic content risk practices for generative AI, and what are the perceived opportunities for improvement of the AI RMF in this context?

Annex - Methodology for Phase 1

A mix-method research methodology was employed, incorporating a combination of qualitative and quantitative methods. We collected data from different sources: desk research, interviews, surveys, case studies, and two Deep Dive online workshops. This mixed-method approach allowed us to triangulate the data and address the research questions for Phase 1 from various perspectives:

Qualitative and Quantitative Methods

Intake Discussions

Individual discussions with participating companies to understand their generative AI adoption, AI risk management practices, and specific interests and practices in AI red-teaming and managing synthetic content risk.

Surveys

Participants joining the program were asked about the sector in which they operate, their levels of familiarity with the NIST AI RMF, and other questions. The first phase of the program encompassed additionally, a short online survey aimed at assessing generative AI adoption, red-teaming/synthetic content usage/practices, and drivers/barriers to adoption across a broad range of organizations. This provided initial quantitative data on prevalence and trends.

Case studies

Participating companies were asked to provide detailed case studies, outlining their motivations for implementing red-teaming/synthetic content risk, specific practices employed, challenges encountered, lessons learned, and recommendations. Case studies offered in-depth understanding of specific implementation experiences and challenges, unveiling nuances beyond broad surveys. Five companies provided case studies with relevant experiences, providing detailed insights into motivations, specific practices, challenges, lessons learned, and recommendations related to red-teaming, and six contributed case studies on synthetic content risk management.

Deep dive workshops

We conducted two interactive online workshops focused on AI red-teaming and synthetic content risk, featuring expert presentations, group discussions, and case study sharing. This fostered deep exploration of current practices and challenges, facilitated by expert insights and peer-to-peer learning, capturing insights not readily captured through surveys.

Limitations and Considerations

The mixed-methods approach proposed for this study is well-suited to address the research questions and objectives. However, the limitations of the methodology should be carefully considered when interpreting the findings of this report.

Self-reported data

Reliance on self-reported information introduces potential bias, requiring cautious interpretation.

Limited sample size

While representative of diverse industries, the sample size may not capture all industry nuances or emerging practices.

Temporal scope

The research captured a specific point in time (from February 2024 to March 2024), and practices may evolve over time.

These limitations necessitate careful interpretation of findings. Triangulation of data from multiple sources and methods mitigates potential biases. While not generalizable to the entire population, the research provides valuable insights and trends within the participating organizations. Future research can expand the scope and address emerging practices.