National Institute of Standards and Technology
Response to Request for Information:
NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence
88 Fed. Reg. 88368 (Dec. 21, 2023)
Docket No. NIST–2023–0309
February 2, 2024

The National Institute of Standards and Technology's (NIST's) request for information on fulfilling its obligations under the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI EO) builds on its AI Risk Management Framework (AI RMF) and other vital efforts to build a robust and flexible set of guidelines, standards, and best practices for trustworthy AI.[1] Google appreciates the opportunity to provide our views, which are informed by our extensive collaboration with policymakers, industry, and civil society, as well as years of work developing and implementing our AI Principles and risk assessment framework.

Executive Summary

Our mission is to organize the world's information and make it universally accessible and useful. Making AI helpful for everyone is crucial to delivering on this mission and improving lives everywhere. We're encouraged to see governments around the world call for ongoing transparency into AI governance processes and AI models' capabilities and limitations.

We've long said that harnessing AI's enormous potential will take the right policy frameworks. We encourage NIST to address generative AI (GAI) with a practical approach that lays the foundation for balanced international standards. Clear and consistent rules can accelerate innovation and give more people access to AI's life-changing benefits.

Developing a GAI Companion Resource. We support NIST's development of an AI RMF companion resource for GAI, which will help enterprises develop and deploy GAI responsibly and manage risks specific to GAI. Consistent with the AI RMF, the GAI companion document should reflect a risk-based approach to developing and deploying GAI tools. We also recommend that the guidance endorse three essential practices for developing responsible GAI:

---

[1] National Institute of Standards and Technology, Commerce Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence, 88 Fed. Reg. 88363 (Dec. 21, 2023); Biden Administration's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, EO 14110 (Oct. 30, 2023).

Designing for Responsibility – Identify, document, and mitigate potential harms (such as unfair bias in AI model outputs) within a system;

Adversarial Testing – Systematically evaluate systems by providing malicious or inadvertently harmful prompts across a range of scenarios; and

Communicating Simple, Helpful Explanations – (1) Where feasible, make it clear to users when and how GAI is used, (2) show users how to offer feedback, and (3) explain how outputs are generated.

This resource should also identify transparency tools used for GAI to help AI developers, deployers, and users consider and maintain best practices. For example, the companion document could promote increased use of transparency artifacts, such as model cards and data cards, which can help developers present information about their models and use cases, as well as provide guidance that would ensure more consistency in how information in these reports is presented. The companion document should also identify common risks associated with GAI and the types of mitigations and interventions that developers and deployers might consider in response.

Creating Benchmarks for Assessing AI Capabilities and Risks. We believe that NIST can make a valuable contribution to responsible AI by identifying benchmarks for assessing AI capabilities and risks. In particular, we suggest that NIST:

- Maintain a flexible risk-based approach, consistent with the AI RMF, in selecting benchmarks, such as providing suites of potential benchmark options that organizations can choose from based on model characteristics and use case;
- Involve a wide array of stakeholders in developing performance standards;
- Define a process to keep benchmarks up-to-date with evolving definitions of harms and risks;
- Encourage AI developers and deployers to incorporate diverse perspectives in evaluating their AI tools;
- Support benchmarking that is accessible to non-technical audiences; and
- Align guidance with emerging global standards development efforts, like those by the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

Establishing Flexible Red-Teaming Guidelines. NIST can help developers by providing more rigor and clarity around AI red-teaming (including by clearly defining AI red-teaming and related terms), but it should keep in mind that red-teaming remains an active and open research space with new methods and tools continuously being developed. NIST should also clarify that there are different types of AI red-teaming with different goals and different implications for issues, such as when and what results to report.

NIST should develop high-level best practices for red-teaming rather than prescribing very detailed procedures—both because red-teaming is inherently exploratory and since best practices will change over time as attackers and defenders innovate. We urge NIST to focus on developing guidelines on (1) how to cultivate the appropriate depth and breadth of expertise, (2) when to use red-teaming in the life cycle of AI systems, and (3) when and how to publish red-teaming findings. Red-teaming norms from the cybersecurity context should also be incorporated into AI red-teaming guidance. In addition, we recommend that NIST clarify that AI risk-management recommendations, including red-teaming, should be applied at different levels as well as at different points in the product development life cycle. NIST's guidance might define the major milestones where red-teaming is recommended, and what specific type of red-teaming is appropriate at that time.

Mitigating Synthetic Content Risks. Current technical approaches to labeling AI-generated content may help mitigate some of the misinformation and other risks associated with synthetic media. For example, SynthID, a tool in beta from Google DeepMind, is an early and promising technique for embedding a digital watermark into AI-generated images and audio. However, all current technical approaches have limitations, and the most effective mitigation approaches will involve close collaboration among ecosystem stakeholders as well as triangulation using various techniques, such as information sharing and user education. Policymakers can leverage lessons learned from the cooperative frameworks already being established by the AI industry in partnership with stakeholders in civil society, media, journalism, and academia, such as the Partnership on AI's Synthetic Media Framework.

NIST guidance could also encourage GAI deployers and users to include disclosures for synthetic content that are tailored to the context and risk level of the content. For example, we were the first tech company to require election advertisers to prominently disclose when their ads include realistic synthetic content, which differs from the approach we have articulated in other contexts, given the specific risks in that area. The value of disclosure depends on various factors, including the type and degree of alteration, potential harm, and intent of the content creator. Background changes, sizing, or lighting enhancements in an image, for example, are typically innocuous. We also recommend that GAI developers restrict the output of particularly dangerous synthetic content.

Advancing Global Standardization. As governments everywhere begin to develop new AI-focused regulation, and US businesses of all sizes seek to incorporate GAI into their goods and services that they seek to export globally, it is critical that we accelerate work on developing internationally aligned, consensus-based AI standards. We urge NIST and US policymakers to continue to engage in these efforts. In particular, NIST should incorporate or reference new or in-development international AI standards into its work, as well as encourage other government agencies to reference these

standards where appropriate. For example, ISO 42001 establishes a management framework for organizations involved in developing, providing, or using AI-based products or services. We also urge NIST to encourage AI stakeholders to participate in ongoing global consensus-building efforts around AI safety, such as those organized by the G7 and the Organisation for Economic Co-operation and Development (OECD), as well as engage directly with these initiatives. These efforts could complement international technical standards by, for example, providing canonical datasets, evaluation methods, and benchmarks for evaluating AI systems.

* * * * * * * *

Introduction: About Google's Efforts to Develop and Deploy AI Responsibly

We advocate a holistic approach to AI safety, security, and privacy, which recognizes that AI models are deployed as parts of broader systems, and different types of testing and transparency are needed for different types of AI systems. Depending on the use case, existing standards and frameworks can guide efforts for building safe and secure technology and help facilitate the safe integration of AI. For example, in the security context, AI developers should incorporate NIST's Secure Software Development Framework and Cybersecurity Framework, which cover broad enterprise and product security best practices.[2] NIST's guidance should likewise encourage AI developers and deployers to fully leverage existing tools and guidance materials developed by NIST and other standards organizations.

Developing and Deploying Gemini: A Case Study

As NIST considers how to develop guidance into GAI best practices, our approach to the Gemini family of AI models provides an informative example of how to implement responsible AI practices in the real world. The Gemini family of models includes our most capable and general model yet, built from the ground up with multimodal capabilities and demonstrating state-of-the-art performance across many leading benchmarks.[3] We've employed the most comprehensive safety evaluations of any Google AI model to date, including for bias and toxicity. We've conducted novel research into potential risk areas[4] (such as cyber-offense, persuasion, and autonomy) and have applied Google Research's best-in-class adversarial testing techniques[5] (such as scaled adversarial data generation, rater diversity, and automated test set evaluation) to help identify critical safety issues ahead of Gemini's deployment.

---

[2] *Secure Software Development Framework (SSDF)*, NIST (last updated Jan. 10, 2023).
[3] *Gemini: A Family of Highly Capable Multimodal Models*, Google (Dec. 6, 2023).
[4] *An early warning system for novel AI risks*, Google DeepMind (May 25, 2023).
[5] *Responsible AI at Google Research: Adversarial testing for generative AI safety*, Google Research (Nov. 16, 2023).

To identify gaps in our internal evaluation approach, we've worked with a diverse group of external experts and partners to stress-test our models across a range of issues. For example, to diagnose content safety issues during Gemini's training phases and ensure its output follows our policies, we've used benchmarks, such as Real Toxicity Prompts,[6] a set of 100,000 prompts with varying degrees of toxicity pulled from the web developed by experts at the Allen Institute for AI. Further details on this work are coming soon.

To limit the risk of harm when deploying Gemini, we built dedicated safety classifiers to identify, label, and sort harmful content (such as depictions of violence or negative stereotypes).[7] Combined with robust filters, this layered approach helps make Gemini safer and more inclusive. We're also continuing to address known challenges for models, such as factuality, grounding, attribution, and corroboration.[8]

Responsibility and safety will always be central to the development and deployment of our models. We have identified at least four basic scenarios[9] of how an organization might be participating in the AI ecosystem that require different risk management strategies. A key difference between these four scenarios is the level of direct control an organization has over the AI model, as compared to what is outsourced to an external provider. Advancing responsibility and safety requires collaboration between government and other stakeholders, including by developing responsible AI standards and benchmarks through organizations like NIST.

I. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

Developing an RMF companion resource for GAI will allow NIST to address important safety issues while continuing to provide the flexibility needed for safe new technologies to flourish. After the release of the first version of AI RMF, Google DeepMind operationalized the AI RMF Playbook and conducted a gap analysis of our responsible AI approach to see if actions under the four AI RMF functions were relevant to our work; which teams were responsible; and if we already had a process in place and if not, if one should be created.[10] We shared with NIST a template to do this gap analysis.

---

[6] Sam Gehman et al., *Real Toxicity Prompts*, Allen Institute for AI (2020).
[7] Sundar Pichai & Demis Hassabis, *Introducing Gemini: our largest and most capable AI model*, Google Blog (Dec. 6, 2023).
[8] *See, e.g.*, *Data Commons is using AI to make the world's public data more accessible and helpful*, Google Blog (Sept. 13, 2023).
[9] *From turnkey to custom: Tailor your AI risk governance to help build confidence*, Google Cloud Blog (Oct. 17, 2023).
[10] *NIST AI RMF Playbook*, NIST.

In developing a GAI supplement to the AI RMF, NIST should tailor its general AI safety approach to GAI-specific concerns and provide examples of specific mitigations and interventions applicable to GAI risks. Given GAI's rapid pace of evolution and adoption, and that fundamental standards are still being developed, NIST should maintain as much flexibility as possible in any guidelines, recommendations, and benchmarks for assessing AI capabilities and ensure that a wide array of stakeholder views are incorporated, just as it did with the Cybersecurity Framework, which has been used to secure critical infrastructure and valuable financial, health, and national security information.[11] As discussed below, red-teaming is integral to building a complete AI safety regime, but it is but one of several techniques relevant to model testing and evaluation. NIST is well-positioned to advise developers and deployers on building their red-teaming capabilities and implementing best practices.

A. Suggestions for a GAI Companion Resource to NIST's AI RMF

Our approach to incorporating GAI into a wide variety of products may provide an instructive example for others.[12] This begins with cross-company, pre-launch process assessments of early product designs against known legal requirements, emerging legislation, standards, and internal AI Principles. From there, teams apply technical or policy mitigations and guardrails, such as filtering training datasets, additional safety filters, or relevant model refinements. Product teams continue to iteratively refine these tools throughout the launch and operations processes.[13] This approach offers useful insights for supplementing the AI RMF in a way that supports responsible GAI development.

1. Take a Risk-Based Approach to GAI Tools

An AI RMF supplement on GAI should reflect a risk-based approach to developing and deploying GAI tools. Responsible innovation requires balancing the ethical risks of new tools with the social benefits of using them.[14] For example, in developing technology that generates photorealistic images of people, we weighed the serious risks of deepfakes and misinformation against the societal benefits of enabling small businesses and creators to make high-quality content that will help grow their

---

[11] The Cybersecurity Framework assists software developers in decreasing vulnerabilities in software releases, minimizing the potential consequences arising from the exploitation of undetected or unresolved vulnerabilities.

[12] *See* Google, *2023 AI Principles Progress Update* at 8-17 (2023).

[13] We continue to scale this approach, doubling to more than 500 AI Principles reviews in 2023, which mainly focused on implementing GAI research models into products, services, and features. *Id.* at 10.

[14] Researchers at Google DeepMind recently published a paper that proposes a three-layered framework for evaluating the social and ethical risks of AI systems. This framework includes evaluations of AI system capability, human interaction, and systemic impacts. Laura Weidinger et al., *Sociotechnical Safety Evaluation of Generative AI Systems* (2023); *see also* Laura Weidinger & William Isaac, *Evaluating Social and Ethical Risks from Generative AI*, Google DeepMind (Oct. 19, 2023).

businesses and contribute to their communities. We developed an approach that makes GAI image technology available in some of our products, subject to strict testing and clear guardrails, such as the use of safety classifiers and filters.

For the supplement, we suggest that NIST promote an approach similar to our risk assessment framework, which seeks to identify, measure, and analyze risks beginning with product conception and continuing throughout development.[15] Our framework allows us to assess potential harms while accounting for myriad possible impacts, including unfair biases and stereotypes, poor product experiences, and information gaps or "data voids."[16] We then map these risks to mitigations and interventions, drawing on best practices from our cross-company enterprise risk management.[17]

To enhance this approach, NIST should also consider including in its GAI supplement methods for AI developers to anticipate and respond to new model capabilities and risks. For example, researchers from Google DeepMind and other organizations proposed a framework for using model evaluation to uncover "dangerous capabilities," including threats to security, the ability to exert influence, or the ability to evade oversight.[18] This framework also includes approaches to measuring model alignment by examining its tendency to apply those harmful capabilities. Alignment evaluations should confirm that the model behaves as intended across a wide range of scenarios and should, where possible, provide effective specifications of model breakpoints or measures of the magnitude of deviation from expected behavior.

2. Tailor Policies and Practices for Responsible GAI Development

Given the wide range of outputs GAI can provide, NIST should also supplement the AI RMF by encouraging GAI developers to implement processes for identifying harmful content and limiting its production. Doing so will help ensure that AI developers are taking the precautions necessary to produce GAI in a way that benefits the public.

To this end, Google's internal AI safety policy is a useful starting point because it proactively identifies a number of known harms associated with GAI outputs. Our policy builds on our extensive experience with harm mitigation, research, record of prioritizing product safety,[19] and commitment to product inclusion and equity.[20] It

---

[15] *See* *2023 AI Principles Progress Update* at 10-11.
[16] Google, *A New Way to Search with Generative AI: An Overview of SGE* (2023).
[17] We conduct AI Principles reviews for all generative AI projects, with particular focus on the following areas: government-related; social impact; recommendation, personalization, and ranking systems; critical technology infrastructure; environmental sustainability; health, fitness, and well-being; finance, education, and employment; surveillance and/or biometrics; and ambient computing, affective technology, and wearables.
[18] Toby Shevlane et al., *Model Evaluation for Extreme Risks* (2023); *see also* Toby Shevlane, *An Early Warning System for Novel AI Risks*, Google DeepMind (May 25, 2023).
[19] *See* *Every Google Product Is Designed for Safety*, Google.
[20] *See* *Building for Everyone, with Everyone*, Google.

states that GAI products must not create harmful content (such as child sexual abuse and exploitation, hate speech, harassment, violence and gore, or obscenity and profanity); dangerous content (i.e., content that facilitates, promotes, or enables access to harmful goods, services, and activities); or malicious content (such as spam or phishing).[21] It also targets the harms caused by misinformation or unfair bias.[22]

This safety policy, which serves as a uniform policy baseline for all GAI products and modalities, provides us with a foundation to implement and improve our AI safety practices. For example, Google uses this policy to help guide product teams and has established a framework to define the types of harmful content that we do not permit our models to generate.[23] We allow customers to adjust a limited set of our safety filters to support their use cases. The framework also guides how we protect personal data, such as health information.

Building on this policy, Google recommends that any supplemental NIST guidance includes these three essential practices for developing responsible GAI.

> **Designing for Responsibility**: This proactive approach focuses on identifying and mitigating potential harms, such as unfair bias in AI model outputs, within a product. These harms can be mitigated at various stages during product development, including by using responsible datasets, classifiers and filters, and in-model mitigations, such as fine-tuning, reasoning, few-shot prompting, data augmentation, and controlled decoding. To support these efforts for stakeholders of all sizes, Google makes available APIs, toolkits, and other resources, such as its 2023 model card updates and collaboration extensions with public compute.[24]
>
> **Adversarial Testing**: Models should be systematically evaluated by providing malicious or inadvertently harmful prompts across a range of scenarios to identify and mitigate potential safety and fairness risks. We conduct such testing before major model and product launches, including our Gemini family of models.[25]
>
> **Communicating Simple, Helpful Explanations**: This involves (1) Where feasible, making it clear to users when and how GAI is used, (2) encouraging users to offer feedback on model safety and behavior, and (3) explaining how outputs are generated by educating users on how they are in control as they use AI products and services. As noted below, this also includes documentation that

[21] *See 2023 AI Principles Progress Update* at 11.
[22] Pandu Nayak, *New Ways We're Helping You Find High-Quality Information*, Google Blog (Aug. 11, 2022).
[23] *2023 AI Principles Progress Update* at 11.
[24] *See, e.g.*, *AI APIs for Google Cloud*, Google Cloud; *Free AI Tools from Google Cloud*, Google Cloud; *The Value of a Shared Understanding of AI Models*, Google Cloud.
[25] *See Gemini: A Family of Highly Capable Multimodal Models* at 20.

makes public essential information based on our internal documentation of safety and other model evaluation details and that can offer guidance for AI researchers, deployers, and developers on the responsible use of the model.

The supplemental guidance should also clarify best practices around transparency. Transparency is vital to responsible AI development and deployment—and guards against harms—because it allows the public to make informed comparisons between models and equips them to recognize problems. That is why Google has developed template documentation tools known as data[26] and model cards,[27] which are used to simplify and standardize information about an AI model or its underlying datasets. Transparent documentation for AI may also involve releasing technical reports or other artifacts that appropriately make additional information public.

Google is also piloting a transparency artifact aimed specifically at integrating GAI models into AI-powered systems: a "generative AI system card."[28] Our first version is intended to provide structured, accessible information for non-technical audiences, such as third-party auditors, policymakers, journalists, enterprise clients, users, and clients and advertisers. The cards offer an overview of the capabilities and limitations of a GAI model when integrated into a larger system that people interact with as a product or service.

More generally, Google supports increased standardization across GAI transparency artifacts, which will help stakeholders compare models and tools and improve public understanding. Our model card toolkit helps model developers prepare these artifacts and provides a roadmap detailing how transparency artifacts could be standardized. Google is also a member of the Partnership on AI's Safe Foundation Model Deployment working group, which is examining a wide range of issues including transparency reporting best practices.[29]

---

[26] Data cards are a dataset documentation framework aimed at increasing transparency across dataset life cycles. They provide structured summaries of ML datasets with explanations of processes and rationales that shape the data and describe how the data may be used to train or evaluate models. *The Data Cards Playbook: A Toolkit for Transparency in Dataset Documentation*, Google AI Blog (Nov. 17, 2022). At a minimum, data cards include the following: (1) upstream sources, (2) data collection and annotation methods, (3) training and evaluation methods, (4) intended use(s), and (5) decisions affecting model performance.

[27] Model cards are short documents accompanying trained machine learning models that typically include information, such as the model's intended use case, the data used to train the model, the model's performance on different metrics, any known biases or limitations of the model, and any potential risks or unintended consequences that could arise from its use. Model cards can also include information about the model's training and evaluation processes and how the model can be deployed and integrated into different applications.

[28] *See 2023 AI Principles Progress Update*, app'x (2023) (example card documenting the December 2023 update of Bard with specifically tuned Gemini Pro).

[29] *PAI's Deployment Guidance for Foundation Model Safety*, Partnership on AI.

Finally, AI providers also need flexibility to protect personal, confidential, and competitively sensitive information,[30] as well as information that, if publicly released, might compromise the safety of AI systems. We recommend that NIST's guidance acknowledge these considerations.

3. Include Common GAI Interventions and Mitigations

We also recommend that NIST identify common risks associated with GAI and the categories of mitigations and interventions that developers and deployers might consider in response. This will enhance NIST's efforts to support responsible innovation.

To illustrate, Google has invested in these kinds of resources to support its own work, increase transparency, and lower barriers to entry. For example, we provide self-service guides that inform developers about methods for addressing GAI risk.[31] We also catalog patterns of GAI risks, including "hallucinations" and model outputs that reflect or reinforce unfair biases or outputs that are extremely similar to or indistinguishable from those created by humans. NIST may wish to include the following AI safety interventions that Google has found useful and continues to use. These interventions fall into four categories:

> **Technical and tooling interventions**, including model monitoring suites[32]; tools for identifying AI-generated content (such as watermarking)[33]; adversarial testing; diverse sets of users, scenarios, and sources considered in data, training, testing, and launch; privacy-preserving algorithms and training techniques[34]; safety guardrails (such as filters, classifiers); built-in model mitigations (such as fine-tuning, reinforcement learning, and advanced capabilities to steer model output to more responsible outcomes);
>
> **Policy restrictions**, including prohibited use policies; policy enforcement and escalation pathways;
>
> **Documentation**, including transparency artifacts (such as model, data, or system cards); disclosures/disclaimers that explain how the model can be used and its limitations (such as a technical report); responsible AI guides; and

---

[30] Google understands that it will be appropriate in some cases to provide policymakers and government officials with additional information, as contemplated under President Biden's AI Executive Order, and we fully support enhanced disclosure of such information, provided appropriate protections for sensitive or confidential information are used.

[31] *See* *Responsible AI Practices*, Google.

[32] *See, e.g.*, *Introduction to Vertex AI Model Monitoring*, Google Cloud.

[33] *See, e.g.*, *Identifying AI-generated images with SynthID*, Google DeepMind (Aug. 29, 2023).

[34] *See, e.g.*, *How Sensitive Data Protection can help secure generative AI workloads*, Google Cloud Blog (Oct. 4, 2023).

Feedback mechanisms, including agile feedback and reporting mechanisms and remediation pathways (such as an appeals process).

B. Creating Guidance and Benchmarks for Evaluating and Auditing AI Capabilities

Google applauds NIST's aim to establish guidance and benchmarks for evaluating and auditing AI capabilities. It is essential for government and industry to collaborate on research and identify canonical methods for evaluating AI capabilities to inform internal AI governance mechanisms as well as support external validation and auditing of AI systems. Below, we offer some recommendations that might help inform NIST's work in this space.

First, we again urge NIST to maintain a flexible approach, consistent with the AI RMF, in light of the rapid pace of technological change, emerging model capabilities, and the nascent state of benchmark research. For example, we encourage NIST to recommend a variety of possible evaluations for specific model capabilities so that model developers can select the evaluations that are most appropriate. Flexibility is also important for incorporating the diverse array of interests, stakeholders, and policy tradeoffs that AI raises. This approach will ensure that the public benefits from transformative AI technologies while also ensuring that these benefits accrue on an equitable basis and that the public is protected from AI harms.

Second, we encourage NIST to involve a wide array of stakeholders in developing standards and to continue seeking input to ensure that any guidance, benchmarks, or recommendations reflect the evolving state of AI technology. We have supported, for example, MLCommons' proposal to utilize a multistakeholder process for selecting tests and grouping them into subsets to measure safety for particular AI use cases, and we are supporting the recently launched MLCommons Working Group to develop and update standard safety benchmarks.[35]

Third, NIST should encourage AI developers and deployers to incorporate diverse perspectives in evaluating AI. In particular, public and private organizations deploying AI technology should engage evaluators with a range of social, ethical, and technical expertise at each stage of an AI system's development and deployment—including problem and governance formation, dataset curation, model development, creation of test-evaluate-validate-verify (TEVV) models, and deployment milestones.[36]

---

[35] *Supporting Benchmarks for AI safety with MLCommons*, Google Research (Oct. 26, 2023).
[36] *See* Sejal Goud et al., *A Blueprint for Equitable AI*, Aspen Institute Science & Society Program at 7 (2023).

Fourth, NIST should support benchmarking that is accessible to non-technical audiences. This is the approach taken by MLCommons.[37] Making information accessible will promote transparency and lower barriers to entry, sparking innovation.

Fifth, as noted below, NIST should, wherever possible and appropriate, align its guidance with global standards development efforts, like those by the ISO and IEC.[38] We have supported efforts by ISO as well as other organizations working to create AI industry benchmarks and evaluations, such as those from the MLCommons[39] and the Frontier Model Forum. Coordinating with these efforts will reduce global compliance burdens, especially for small companies, while allowing NIST to focus its limited resources on the highest-priority areas.

C. Providing Flexible Guidance on Red-Teaming

Google supports NIST's approach to creating red-teaming guidelines for improved AI safety and security, but we are concerned about a lack of definitional clarity about what red-teaming is and how to deploy it most productively. The AI EO defines "AI red-teaming" 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."[40] Similarly, one of the Frontier Model Forum's[41] earliest efforts was defining what constitutes "red-teaming" for frontier AI models. The Forum members agreed to define red-teaming "as a structured process for probing AI systems and products for the identification of harmful capabilities, outputs, or infrastructural threats."[42] In practice, however, it seems that "red-teaming" is often used as a catch-all, encompassing a broad sweep of AI safety testing practices, which is confusing and potentially counter-productive.

---

[37] *Supporting Benchmarks for AI safety with MLCommons*, Google Research (Oct. 26, 2023).
[38] The ISO recently published ISO 42001, which sets standards for AI management systems, helping to create a "structured way to manage risks and opportunities associated with AI, balancing innovation with governance." ISO 42001:2023. The ISO and IEC are also considering standards for AI system impact, ISO 42005, and auditing and certification of AI management systems, ISO 42006.
[39] *Supporting Benchmarks for AI safety with MLCommons*, Google Research (Oct. 26, 2023).
[40] Biden Administration's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, EO 14110 (Oct. 30, 2023). The AI EO also notes that "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."
[41] In 2023, Google, Anthropic, Microsoft, and OpenAI formed the Frontier Model Forum, a new industry body focused on ensuring safe and responsible development of frontier AI models. The Frontier Model Forum will draw on the technical and operational expertise of its member companies to benefit the entire AI ecosystem, such as through advancing technical evaluations and benchmarks, and developing a public library of solutions to support industry best practices and standards. *See generally Frontier Model Forum*.
[42] *Frontier Model Forum: What is Red Teaming?*.

Our AI red teams focus on testing AI models and products for a wide variety of risks, including security, privacy, and abuse risks.[43] Red-teaming is a critical capability that we deploy to support our AI Principles and security frameworks,[44] and we believe that our experiences can inform NIST's guidance and recommendations. In our view, however, red-teaming is only one of several types of evaluation and testing processes that we, and other AI developers, have noted are important for developing and deploying powerful AI systems safely. Red-teaming should be considered a necessary but not sufficient component in any AI safety testing and evaluation process. NIST can help developers by providing more rigor and clarity around AI red-teaming, including by clearly defining key terminology in its guidance. We encourage NIST to focus on the probative and exploratory nature of red-teaming in order to provide useful guidance to practitioners; red-teaming results are often used to inform other safety testing efforts, such as the use of specific quantitative evaluations and benchmarks.

We applaud NIST's ongoing work to establish a taxonomy and terminology for machine learning,[45] and encourage NIST to provide additional clarity around these key terms: adversarial simulation (i.e., how red-teaming is used in the security space); adversarial testing in different areas (e.g., bias, fairness, toxicity, persuasion); and dangerous capabilities testing (e.g., weapons development, offensive cybersecurity capabilities, and persuasion).

Red-teaming remains an active and open research space, and new methods and tools are continually being developed. As such, rather than highly prescriptive procedures for how to engage in red-teaming that might stifle innovation in this space, we encourage NIST to focus on developing high-level best practices around red-teaming, including (1) how to cultivate the appropriate depth and breadth of expertise, (2) when to utilize red-teaming in the life cycle of AI systems,[46] and (3) when and how to publish red-teaming findings.

For instance, while we believe that AI developers should provide information on the types of red-teaming or other safety testing they have conducted, our experience shows that it is not always appropriate to share all results from red-teaming exercises. Broadly, we believe that NIST should focus on reporting results that directly relate to model capabilities and safety. Moreover, many red-teaming results are highly sensitive;

---

[43] Several teams within Google are involved within the broad concept of "red-teaming." For example, adversarial prompting mostly falls outside the scope of machine learning red-teaming but constitutes a part of the broad "red-teaming" array within Google.

[44] Google, *Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems* (July 2023).

[45] Apostol Vassilev, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2023 (Jan. 2024).

[46] NIST's guidance should note that red-teaming should be done at different (1) "levels" (e.g., at the model level and at the application level (where the model is integrated into a product)) and (2) points in the product development life cycle. The guidance should also provide suggestions on red-teaming activities at major milestones within a model's or product's life cycle.

in some cases, it may be appropriate to share findings only in aggregate form or to share information on certain sensitive topics—such as information related to nuclear or biological weapons—with governments but not publicly. In contrast, other areas like fairness, bias, and toxicity are highly relevant to the public, so publishing benchmarks or scorecards on those topics may be appropriate. NIST should provide flexibility concerning the reporting of red-teaming testing to protect confidential and sensitive information, minimize the ability to reverse engineer AI systems, and prevent publication of information that might help users circumvent safety mitigations.

AI red-teaming is also not a single practice; rather, it encompasses a range of testing practices and procedures, each having unique implications from a system safety or transparency perspective. For example, AI developers will often engage in early exploratory red-teaming of new models to understand model capabilities and limitations. Although these findings are helpful in informing subsequent evaluation, red-teaming, and mitigation practices, it is unlikely that insight into such provisional results would be of much substantive value to either policymakers or the public.

Another subset of practices involves traditional adversarial testing of models for flaws and vulnerabilities. For example, our Red Team consists of a team of hackers that simulate various threat actors, ranging from nation-states and well-known Advanced Persistent Threat (APT) groups to hacktivists, individual criminals, or even malicious insiders. Over the past decade, we've evolved our approach to translate this concept of red-teaming to the latest innovations in technology, including AI. Our AI Red Team is closely aligned with traditional red teams but also has the necessary AI subject matter expertise to carry out complex technical attacks on AI systems.

A key responsibility of Google's AI Red Team is to take relevant research and adapt it to work against real products and features that use AI to learn about their impact. We leverage attackers' tactics, techniques, and procedures (TTPs) to test a range of system defenses.[47] This includes, for example, prompt attacks, training data extraction, backdooring the model, adversarial examples, data poisoning, and exfiltration, but also covers common cybersecurity issues pertaining to the integration of AI models as software components. To ensure that they are simulating realistic adversary activities, our team leverages the latest insights from world-class Google Threat Intelligence teams like Mandiant and the Threat Analysis Group (TAG), content abuse red-teaming in Trust & Safety, and research into the latest attacks from Google DeepMind.

[47] Google's red-teaming report includes a list of TTPs that we consider most relevant and realistic for real-world adversaries and red-teaming exercises. *Why Red Teams Play a Central Role in Helping Organizations Secure AI Systems*.

We also incorporate principles included within Google's Secure AI Framework (SAIF),[48] which provides several recommendations for assuring the security of AI systems.[49] The SAIF could serve as a conceptual framework that both government and the private sector could use to collaboratively secure AI technology.

NIST should incorporate cybersecurity norms into its approach to this type of red-teaming, for example, by affording model developers an appropriate time period to remedy any identified vulnerabilities before reporting any findings pursuant to testing. Similar to the security space, these vulnerabilities need not be published or reported publicly unless (1) users need to take action to fix the vulnerability (e.g., installing an update) or (2) the vulnerability was maliciously exploited and users or customers were affected.

Yet another type of red-teaming involves adversarial testing of our AI tools for content policy violations and to measure how well a model is following our policy framework and how well our technical safety mitigations and fine-tuning are working. While we generally expect our GAI products to restrict content prohibited in our internal AI safety framework, there are some important exceptions.[50] Given the importance of reducing model hallucinations, biased content, and the generation of potentially dangerous information (including content detailing steps on how users might cause harm to themselves or others, Child Sexual Abuse Material (CSAM), or other explicitly violative content), red-teaming aimed at ensuring content quality is an important, and distinct, input to AI model safety. However, such evaluations also involve more difficult and nuanced judgment calls than traditional cybersecurity-oriented red-teaming. We encourage NIST to tailor its recommendations to account for these factors and give due care (and detailed guidance) to how developers are to balance the significant legal implications and technical limits of adversarial testing in sensitive content domains, such as CSAM or similar content. We also continue to innovate with methods for scaled automated testing using large language model-based auto-raters to enable efficiency and scaling.

---

[48] *Introducing Google's Secure AI Framework*, Google Blog (June 8, 2023). SAIF aims to address the unique risks associated with AI systems, such as unauthorized model theft, contamination of training data, introduction of malicious inputs through prompt injection, and unauthorized extraction of sensitive information from the training data. *Id.*

[49] SAIF recommendations include (1) expanding strong security foundations to the AI ecosystems, including secure-by-default protections, (2) extending detection and response to bring AI into an organization's threat universe, (3) automating defenses to keep pace with new and existing threats, (4) harmonizing platform-level controls to ensure consistent security across organizations, (5) adapting controls to adjust mitigations and create faster feedback loops for AI deployment, and (6) contextualizing AI system risks in surrounding business processes.

[50] Similar to other Google products, for example, featured snippets on Search, we make an exception when there is an educational, documentary, scientific, or artistic benefit to showing or translating content that might otherwise be perceived as offensive within these specific, beneficial contexts, as we do within the Bard experience.

We believe that public red-teaming is an important supplement to other types of red-teaming and testing practices and encourage NIST to develop guidance on best practices for conducting such red-teaming, including clarifying that public red-teaming is appropriate for some content domains (such as testing models for bias) but not for others (such as CSAM) and methods to limit harm to participants.

We also recommend that NIST clarify that AI risk-management recommendations should be applied at different levels, such as at the model level and the application level (where the model is integrated into a product), as well as at different points in the product development life cycle. NIST's guidance might define the major milestones where red-teaming is recommended and what specific type of red-teaming is appropriate at that time. For example, it may be appropriate to conduct additional red-teaming after a model has been fine-tuned for a particular use case. External red-teaming should only be required or recommended where it is necessary and technologically feasible. More generally, safety testing should reflect both the potential risk the application presents and existing capabilities—including recognizing the role of internal red-teaming.

II. Reducing Synthetic Content Risks

Google applauds NIST's efforts to reduce the risks associated with synthetic content. This work is essential to addressing the harms we are already seeing in the real world and bolstering public confidence in AI technology. AI's ability to produce synthetic content has many useful applications, such as opening new possibilities to those affected by speech or reading impairments[51] and providing new creative opportunities for artists and movie studios.[52] But it can also be used for malicious purposes, such as disinformation campaigns.

Synthetic content is an emerging field of study. As NIST explores methods for authenticating, labeling, detecting, testing, and auditing synthetic content, it should recognize that current technical approaches have limitations. For example, classifiers have demonstrated only limited success to date in accurately determining whether images have been generated using GAI tools.[53] While research into synthetic data detection tools remains ongoing, NIST should maintain a flexible approach that does not over-rely on a particular technology or technical solution or assume that any one solution alone will be sufficient.

---

[51] *See, e.g.*, *Project Relate*, Google Research (describing an app created to help people with non-standard speech to make their voices heard).
[52] *See* *Made On YouTube: Empowering anyone to Create on YouTube*, YouTube Blog (Sept. 21, 2023); *An early look at the possibilities as we experiment with AI and Music*, YouTube Blog (Nov. 16, 2023); *Our principles for partnering with the music industry on AI technology*, YouTube Blog (Aug. 21, 2023).
[53] *See* Jordan J. Bird & Ahmad Lotfi, *CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images* (Mar. 24, 2023).

Rather than focus exclusively on a limited set of imperfect technical tools available to AI developers, NIST should recognize that risks associated with synthetic content involve a wide range of actors and that the most effective approaches to reducing those risks will involve close collaboration among ecosystem stakeholders. NIST should proactively work with a wide range of stakeholders from across society to triangulate potential approaches to managing risks from synthetic content. While some of these approaches may rely on new or existing technical methods to reduce the risks of synthetic content, such as increasingly sophisticated tools to evaluate content and confirm its provenance, other productive approaches may involve collaboration and information sharing among AI developers, AI deployers, and media platforms to help identify and limit the spread of synthetic media as well as increased consumer education on AI and digital media issues.

To enhance collaboration, NIST should support and incorporate lessons from the cooperative frameworks already being established by the AI industry in partnership with civil society, media/journalism, and academia. For example, Google contributed to the Partnership on AI's development of a Synthetic Media Framework, which aims to foster best practices for developing and sharing AI-generated media.[54] The framework establishes universal principles for participants in the creator economy and offers tailored recommendations for categories of participants, such as research and development focus areas for technology builders.[55]

A. Provenance Methods

For information to be trustworthy, it is essential that readers and viewers be able to identify AI-generated content in relevant contexts. Watermarking can be an extremely useful tool to help reduce risks in these situations by embedding information directly into AI-generated content. Google has begun to build our models to include watermarking and similar capabilities from the start, consistent with our commitments to safe, secure and trustworthy AI.[56] For instance, SynthID, a tool in beta from Google DeepMind, is an early and promising technique for embedding a digital watermark into AI-generated images and audio in a way that is resilient to common editing techniques and transformations.[57] The watermark is imperceptible to humans but detectable for identification. However, NIST should remain mindful of these techniques' limitations, such as their susceptibility to adversarial removal.

---

[54] Partnership on AI, *PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action* (2023).
[55] *Id.* at 4.
[56] The White House, *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI* (July 21, 2023).
[57] Melissa Heikkilä, *Google DeepMind Has Launched a Watermarking Tool for AI-Generated Images*, MIT Tech. Rev. (Aug. 29, 2023); *see also SynthID*, Google DeepMind.

Effective use of metadata can also help curb the risks of synthetic content. Content creators can use metadata to associate context with original files. This, in turn, gives viewers more information about the images they see, such as where the images come from and how they were made. This information can help prevent malicious use of synthetic content or reduce its impacts by keeping people informed about the source of what they are seeing. Given these benefits, Google is ensuring that all our AI-generated images have this kind of metadata.

B. User-Facing Disclosures

Google is a strong supporter of appropriate disclosures for synthetic media. For example, we have developed a robust suite of policies and practices aimed at election integrity, which is a particularly critical context for synthetic media.[58] Google is proud to be a pioneer in this area as the first tech company to require election advertisers to prominently disclose when their ads include realistic synthetic content. We hope that our experience deploying these safeguards is informative for NIST and other stakeholders.[59] Augmenting these protections, Google's "about this result"[60] and "about this image"[61] features help people assess the context and credibility of what they are shown.[62] Moreover, our "double-check" feature enables people to evaluate whether other Internet content confirms English language responses provided by some of our AI tools.[63] We suggest that NIST encourage AI developers and deployers to integrate these types of tools, especially as the 2024 US presidential election unfolds.[64]

It is also crucial that developers include restrictions on synthetic content that is particularly dangerous. For example, Google bans users from generating content

---

[58] *See generally* Leslie Miller, *Supporting the 2024 United States Election*, YouTube: Official Blog (Dec. 19, 2023).
[59] *Political Content - Advertising Policies Help*, Google.
[60] Hema Budaraju, *How We're Responsibly Expanding Access to Generative AI in Search*, Google Blog (Sept. 28, 2023).
[61] Nidhi Hebbar & Christopher Savčak, *3 New Ways to Check Images and Sources Online*, Google Blog (Oct. 25, 2023).
[62] In addition, Google Jigsaw is developing new technology and conducting research to protect open societies and inspire scalable solutions. The Jigsaw team experiments with innovative uses of technology to defend against emerging threats. These experiments help us test new ideas and inform product development. Jigsaw also builds products that help people around the world stay safer online and creates programs to apply its research, technology, and training to critical issues such as election integrity and violence against women online. *See generally* *Jigsaw*, Google.
[63] Yury Pinsky, *Bard Can Now Connect to Your Google Apps and Services*, Google Blog (Sept. 19, 2023).
[64] To be clear, these kinds of external, user-facing disclosure mechanisms make sense in the election context but not necessarily for more generic AI-enhanced content, such as run-of-the-mill ads. Election interference is a particularly acute risk, and overuse of user-facing features could be unhelpful, causing label fatigue and implying the truth of all content without warnings.

related to child sexual abuse, fraud, and terrorism, among other malicious or illicit activities.[65]

However, synthetic media disclosure practices—especially for end users—should be tailored for context and risk level to avoid notification fatigue and poor user experience. For instance, it is less important for users to receive notification of the use of synthetic AI context for creative works. As we have seen with website operators' attempts to comply with the EU's General Data Protection Regulation's website cookie disclosure requirements, overly broad disclosure obligations can be highly disruptive while providing limited safety benefits.

III. Advancing Responsible Global Technical Standards for AI Development

A. Facilitating Hub-and-Spoke Regulatory Models and Supporting the Adoption of Industry-Driven Standards

Global technical standards for responsible AI are important for AI sustainability and development. They reduce global compliance burdens for AI developers and deployers—especially small- and medium-sized businesses—while ensuring a more consistent user experience. As a standards development organization with a wealth of technical expertise, NIST is well-positioned to encourage the development and implementation of global consensus-based standards driven by AI stakeholders, and we encourage NIST to continue to engage in these efforts and reference these standards in its guidelines and documents as appropriate.

Google supports a balanced, risk-based framework for overseeing AI to encourage responsible development and deployment of AI technologies without unduly hampering these tools' broad benefits. We encourage NIST to serve as the interagency "hub" to sector and regulator "spokes"—helping to inform government agencies on relevant standards, best practices, and guidelines for assessing the safety and trustworthiness of AI systems under their jurisdiction. To advance these efforts, we encourage NIST to consider systematizing the AI RMF crosswalk program[66] via open access API and expanding the program to incorporate multistakeholder, international standards. For example, ISO recently published the ISO 42001 standard, which establishes a management framework for organizations involved in developing, providing, or using AI-based products or services. We believe this standard represents a significant advancement in AI risk management, and we've built an AI and advanced technologies governance program that is aligned with the ISO 42001 approach. We encourage NIST to incorporate ISO 42001 into its own recommendations for trustworthy AI and to work closely with interagency stakeholders to incorporate this standard by reference into regulation where appropriate. We encourage NIST to remain

[65] *Generative AI Prohibited Use Policy*, Google (Mar. 14, 2023).
[66] NIST, *Crosswalks to the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (2023).

closely involved with AI standards under development—including ISO 27090, 27091, 42005, and 42006—and to adopt a similar approach to them as ISO 42001.[67]

B. Advancing Globally Harmonized Safety Practices

NIST should also continue to closely monitor and engage with international efforts to develop accepted principles and standards through global consensus-building. AI is inherently a cross-border technology,[68] and we're beginning to see emerging outlines of an international framework for responsible AI innovation.[69]

For example, countries like Australia, Chile, New Zealand, Singapore, and the UK have pioneered new trade agreements that support international alignment of AI frameworks and facilitate the cross-border use of AI technologies. The United Nations also announced an AI advisory group.[70] Additionally, the G7, OECD, ISO, and other international bodies have developed a series of principles, commitments, and standards on AI that can help guide safe, secure, and responsible development.

NIST should encourage AI developers and deployers to participate in such international consensus-building efforts and consortia as well as engage directly with these organizations. For example, Google participates in the Global Partnership on AI, a multistakeholder initiative established by the G7, which aims to bridge the gap between theory and practice on AI by supporting cutting-edge research and applied activities on AI-related priorities. As mentioned above, we are also working closely with non-governmental organizations such as the Partnership on AI, MLCommons, and the Frontier Model Forum to build out industry-wide best practices and standardized testing methods. Continuing to build on these existing and emerging frameworks could complement ISO technical standards by, for example, providing canonical datasets, evaluation methods, and benchmarks for evaluating AI systems. We encourage NIST to partner with these organizations and work to ensure they minimize duplicative work, such as developing guidance that contradicts or replicates issues addressed by an existing or in-development international standard.

Conclusion

Google appreciates this opportunity to comment on how the public and private sectors can collaborate to develop standards and benchmarks for the development and deployment of trustworthy AI.

---

[67] ISO 27090; ISO 27091; ISO 42001:2023; ISO 42005; ISO 42006.

[68] *See, e.g.*, Google, *The AI Opportunity Agenda* ("Finally, because AI is by its nature a cross-border technology, individual policy efforts must be tethered to strong trade and investment policies that support trusted international collaboration on AI, including cross-border data flows essential to AI development and deployment.").

[69] Kent Walker, *A Patchwork of rules and regulations won't cut it for AI*, The Hill (Nov. 05, 2023).

[70] *Id.*