Meta

575 7th St NW
Washington, DC
20004
United States

February 2, 2024

The Honorable Dr. Laurie Locascio
Director
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Subject: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11) - Docket No. 2023-28232

Dear Director Locascio,

Attached please find the comments of Meta Platforms, Inc. in response to the National Institute of Standards and Technology's (NIST's) Request for Information related to the agency's assignments under Sections 4.1, 4.5 and 11 of the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Sincerely,

Brian F. Rice
Vice President, Public Policy
Meta Platforms, Inc.

Executive Summary

Meta welcomes NIST's collaborative approach to its undertakings under the Executive Order and is eager to continue working with NIST on its initiatives, alongside our existing collaborations with the Open Loop program and the AI Safety Consortium.

We believe that an open approach to innovation and research is key to achieving the Executive Order's goals of fostering inclusivity and equality, democratizing access to technology, and promoting safe, secure, and trustworthy technological development. These principles have been guiding our responsible approach to product development and are reflected in our commitment to responsible AI[1].

Our high-level recommendations are as follows:

1. **NIST should leverage existing frameworks:** NIST should build upon the AI RMF, which companies are already using as an important tool. In particular, we encourage NIST to collaborate with all stakeholders to identify and fill any gaps within the framework and to ensure that it provides specific guidance on how to build, fine tune, and deploy generative AI systems safely. Additionally, NIST should leverage ongoing cross-industry and intergovernmental standards-building processes (i.e., Partnership on AI's Synthetic Media Framework and Guidance for Safe Foundation Model Deployment) rather than starting from scratch.

2. **NIST should prioritize evaluation metrics and benchmarks:** NIST's priority should be identifying specific and clear sets of risks unique to generative AI, as well as methodologies and benchmarks to evaluate them, especially focusing on determining which models are "dual-use". Actors across the AI value chain should be provided with "red lines" for clearly delineated outcomes respective to their roles in the value chain, including specific risks for which to test and mitigate, rather than requiring assessments of all conceivable scenarios.

3. **NIST should foster open innovation:** As NIST pursues the Executive Order's core values of promoting innovation, safety, security, competition, and access, it should encourage responsible open innovation and drive consensus towards governance frameworks that facilitate it. These should also take account of the unique needs of research, which strongly relies on open source.

4. **NIST should focus on harmonization of global standards and frameworks**: NIST should leverage its experience to drive standards development, support U.S. leadership in existing standards-building processes and international organizations, and facilitate alignment across the multiple agencies entrusted by the Executive Order. In particular,

[1] https://ai.meta.com/responsible-ai/

NIST should focus on facilitating consistent taxonomies (i.e., metrics, benchmarks, and definitions), including by ensuring the concepts of the AI RMF are mapped to and aligned with the Executive Order, where relevant.

Table of Contents

I. Introduction

Meta Platforms, Inc. welcomes the opportunity to contribute to the work of National Institute of Standards and Technology (NIST) under the *White House Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.*

We have invested in the responsible development of artificial intelligence for more than a decade because we believe that AI has the potential to bring immense benefits to humanity. From improving productivity, to helping reduce healthcare costs and outcomes with new scientific discoveries, to contributing to solving societal problems like climate change, the applications are endless. For instance, Llama 2, our open sourced large language model released in July 2023, has been used to build a recruiting tool for AI-skilled job seekers, to launch a copilot to support scientists in research and development in molecular design, and to boost productivity and collaboration in virtual team meetings.[2] In September, we reported that more than 30 million Llama-based models were downloaded through Hugging Face, and more than 7,000 projects on GitHub mentioned Llama 2.[3] Currently, we count more than 19,000 derivative models built on Llama 2, and more than 5,000 citations of our Llama *2* Research Paper.

As Secretary Blinken observed during the United Nations General Assembly, AI may help advance more than 80% of the UN Development Goals and their targets.[4] The U.S. continues to be a global leader in AI innovation[5] and researchers at Goldman Sachs are estimating AI could bring a potential GDP growth boost of 0.4% points to the U.S.[6] To ensure these positive results, it is imperative that technology is built and deployed responsibly, in the open, and in a way that is grounded in high quality research.

Now more than ever, there is a need for government, industry, academia, and civil society to work together to set common and harmonized AI standards and governance models: codes of practice, standards, and guardrails should be agreed upon consistently around the world.

For this reason, Meta supports the collaborative approach undertaken by the U.S. government to seek consensus around how the most advanced AI should be developed responsibly. In fact, in July 2023 we voluntarily agreed to a set of commitments on AI. These White House commitments are an important first step towards responsible guardrails for the most advanced models and a helpful lead for other governments to follow.[7]

While more universal standards for advanced AI are still being established we are already developing our AI products and services with the same commitments to safety, security, and trust that are at the heart of the Executive Order. Based on lessons learned over the last decade,

[2] See these and more stories here: https://llama.meta.com/community-stories/
[3] https://ai.meta.com/blog/llama-2-updates-connect-2023/
[4] https://www.state.gov/secretary-antony-j-blinken-at-the-ai-for-accelerating-progress-on-sustainable-development-goals-event/
[5] https://hbr.org/2023/12/charting-the-emerging-geography-of-ai
[6] By 2034. https://www.goldmansachs.com/intelligence/pages/ai-may-start-to-boost-us-gdp-in-2027.html
[7] https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

we are building comprehensive safeguards into our AI products from the beginning and, we believe, have perspectives to share on technical feasibility of proposals, state of the art, and the direction of technological advances.

A. The NIST AI Risk Management Framework is an important and relevant tool

NIST's work has been pivotal to the advancement of clear, consistent, and responsible guidelines on responsible AI. Clarity and consistency are particularly important in the fast-developing generative AI space, where extensive work already is underway in many countries and multilateral fora to align on key issues like taxonomies of risks.

NIST has adopted a sensible, technology-neutral approach via its AI Risk Management Framework (AI RMF). This tool has provided tangible and measurable processes to help companies identify risks connected with specific uses of AI and provides companies with important guidance to manage them.

Since its release, the AI RMF has informed Meta's continued efforts to build and mature an end-to-end risk management system for AI.

B. Meta values close collaboration with NIST

Given the importance of the AI RMF and the need to gain experience with it as a tool for responsible development of generative AI, we recently announced our first **Open Loop prototyping program** in the United States, in collaboration with Accenture's Global Responsible AI team. Through the program, we're convening AI industry participants so that they – and the AI ecosystem more generally – can gain practical experience in applying the AI RMF to generative AI. The program will also provide a forum for gathering qualitative feedback to help inform NIST's future guidance on generative AI and further iterations of the AI RMF. In particular, it will be important to understand the experiences of companies seeking to implement the AI RMF for generative AI, including their choices with respect to technical benchmarks and internal standards. It will also be important to consider the technical and qualitative experiences of companies throughout the AI value chain, particularly as they differ with respect to size, maturity level, and role. A focus on different actors in the AI ecosystem can complement the ongoing development of AI RMF "profiles", including for example the one created for general-purpose AI systems by UC Berkeley researchers Tony Barrett, Jessica Newman, and Brandie Nonnecke at UC Berkeley CLTC.[8]

Open Loop is just one example of the importance that Meta attaches to collaborative processes with multiple stakeholders with the aim of establishing effective and implementable best practices for AI risk management.

[8] https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

Meta's plan to participate in the **U.S. AI Safety Institute Consortium** is another example, and we look forward to collaborating with the Consortium to establish a new measurement science that will enable the identification of proven, scalable, and interoperable measurements and methodologies to promote development of trustworthy AI and its responsible use.[9]

Meta looks forward to providing the Consortium with deep expertise in responsible AI development from both a technical and policy perspective. In particular, we look forward to engaging on the following areas outlined in NIST's Consortium Overview: data and data documentation, AI governance, AI safety, trustworthy AI, responsible AI, AI system design and development, AI system deployment, AI red-teaming, test, evaluation, validation and verification methodologies, AI fairness, and AI explainability and interpretability, which strongly align with our priority safety and security work.

We are confident that the Consortium will be an inclusive and efficient way to address the priorities of the Executive Order, gather industry best practices, and help develop consistent standards, and look forward to collaborating with NIST and the Consortium to deliver on that vision.

*   *   *

We organize these comments into three sections, which mirror the three main assignments received by NIST:

Section II covers the development of guidelines and best practices for AI safety and security, highlighting existing challenges in achieving these goals and the steps Meta has taken to support these policy objectives in our AI development.

Section III outlines the investments that Meta has made to consider the benefits and risks of synthetic content. Based on our participation in the Partnership on AI's work on this issue and discussions with other stakeholders, we suggest an approach to allocating responsibilities across the AI value chain and its various actors.

Finally, **Section IV** highlights the impact NIST can have on the development of global standards and the harmonization of global efforts and identifies ways that the U.S. Government can promote strong, harmonized protections for responsible AI development globally.

[9] https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

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

A. NIST should leverage existing frameworks and build upon them

Multiple governmental and industry-wide initiatives are already underway to address AI safety and security. As a result, NIST should take the work that has already been done and build upon it.

i. The White House Commitments and the UK Safety Summit

Last July, Meta joined 6 other leading AI companies in agreeing to the *White House Voluntary Commitments on Safe, Secure, and Trustworthy AI*. At that time,[10] the White House established eight principles focused on ensuring the responsible development of "Frontier AI," including commitments to external red-teaming for specific risks, transparency regarding system capabilities and limitations, provenance and/or watermarking of AI-generated audiovisual content, information-sharing with peer companies, and promotion of "bug bounties" and other mechanisms to identify and report vulnerabilities.

These commitments embody an emerging industry-wide consensus and provide a solid foundation for evolving AI technology, as recognized by the Executive Order's goal of leveraging them as a basis for international frameworks.[11] The *Bletchley Declaration*, which came out of the UK AI Safety Summit,[12] similarly focuses on frontier AI and specifies the need to understand relevant safety risks through a shared, evidence-based approach. Its signatories also set out to build risk-based policies in their respective countries. This is an area where NIST will be pivotal: the AI Safety Institute will serve as the U.S. national center for driving model safety evaluation. This work will be at the basis of international collaboration with similar institutions in partner countries, such as the UK Safety Institute.

ii. UC Berkeley AI Risk-Management Standards Profile & PAI Guidance for Safe Foundation Model Deployment

At an industry level, there are two instruments in particular that we suggest NIST consider in its work. The first is *AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models*,[13] issued by University of California Berkeley AI Research Lab, which provides specific guidance on how to apply the NIST AI RMF to providers of general-purpose AI

[10] July 2023, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

[11] Sec. 11 (ii)

[12] https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[13] https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

systems – which is intended as an umbrella term for foundation models, frontier models, and generative AI. This is a particularly useful framework for NIST to reference as it develops further guidance, particularly the "Govern", "Map", "Manage", and "Measure" functions.

The second is *Guidance for Safe Foundation Model Deployment*,[14] drafted by the Partnership on AI (PAI), an industry group in which Meta actively participates. This represents the most comprehensive and multi-stakeholder AI self-regulatory approach to date, and in addition to aligning with the White House Commitments,[15] it introduces important recommended practices, including for open-sourcing of AI models. Notably, it also touches upon the distribution of responsibility across all actors in the AI value chain and its lifecycle. We suggest NIST leverage this instrument, adapting it as needed.

iii. Other industry-level initiatives

Worthy of mention is also the **AI Alliance**,[16] a membership organization Meta co-founded, which specifically advances open source and open innovation approaches. The AI Alliance seeks to work with global stakeholders to create resources and benchmarks for the responsible and safe development and deployment of AI. In particular, NIST should leverage the Alliance's work to understand ways in which open innovation can not only be encouraged, but even facilitated.

And lastly, Meta also supports and contributes to the AI Safety Working Group[17] at the **MLCommons** non-profit consortium, which aims to establish and maintain a toolkit of globally-viable, standardized metrics and benchmarks for evaluating and measuring the risk of generative AI. This work includes building and maintaining AI tooling infrastructure for benchmarking large language models for safety. In 2024, Meta seeks to open source some of our research outputs, datasets, and other assets in collaboration with MLCommons, which we believe is key to helping the AI community's research efforts in helping to bring the Responsible Use Guide to life around a shared set of evaluations and mitigations.

All of these efforts already constitute a solid foundation. We believe NIST could have the most impact by identifying what unique value they bring, supporting developers in understanding how these various frameworks connect to each other and how to use them, and by helping drive consensus towards what responsible development for AI means. What is needed right now is not new frameworks, but rather alignment on what the most appropriate ones are and guidance on how to implement them.

B. NIST should prioritize metrics and benchmarks for model evaluation

As these examples demonstrate, there is broad agreement about the importance of risk

[14] https://partnershiponai.org/modeldeployment/

[15] In the Commitments from July 2023: https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

[16] https://thealliance.ai/

[17] https://mlcommons.org/working-groups/ai-safety/ai-safety/

assessment and mitigation. However, we do not yet have an agreed upon set of risks, nor are there shared ways to evaluate models against those risks or benchmarks, or to know how one model compares against another. Therefore, as it begins the next phase of its work on AI, we encourage NIST to develop stronger alignment on **model evaluation, benchmarks, and frameworks** to ensure that the ecosystem is able to assess, concretely and consistently, whether models present tangible risks.

Specifically, to address the existing gaps, we encourage NIST to help define a taxonomy of harms and then support the development of metrics to evaluate severity against those harms, the capabilities that might trigger them, and any relevant frameworks and evaluation benchmarks to make these assessments.

i. NIST should help identify and address risks specific to generative AI, focusing on actionable, imminent challenges

While the AI RMF was built with traditional AI in mind,[18] its flexibility makes it suitable for evolving technologies, including generative models. However, generative AI presents new opportunities and, consequently, it might change the nature and scale of existing issues (such as bias or mis/disinformation), or even give rise to unique ones.

To help downstream developers identify, assess and address potential issues as they use our models, and specifically our Llama 2 open sourced large language model, we released a *Responsible Use Guide* that highlights steps a developer should take to use Llama 2 responsibly.[19] These include:

1. **Determining the use case:** the developer should identify the specific purpose for which the model will be used;
2. **Preparing the data**: the developer should consider three main issues: i) *Privacy*: whether datasets include non-publicly available information; ii) *Provenance:* the possible inclusion of content from unknown origins; iii) *Diversity:* ensuring that datasets include representation of different demographic groups; and
3. **Fine-tuning the model:** the developer can adapt the pre-trained model to the specific use case by training it on specialized datasets and introducing additional layers of safety mitigations. In particular, with fine-tuning, the developer can define the policies the model should adhere to, evaluate and/or improve the model's performance, and address any risks arising at an input and output level. In terms of input and output level risk mitigations, the following aspects should be considered:

[18] Traditional AI is capable of analyzing large amounts of data to classify and label content, predict what content users will find most useful and recommend said content to them. Generative AI, on the other hand, emulates "the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content" (Executive Order, Section 3. (p)). While the AI RMF is flexible and should continue to be leveraged, NIST should focus on framing the risks that are specific to generative AI to understand exactly what additional resources, areas and guidance, if any, may be required.

[19] https://ai.meta.com/static-resource/responsible-use-guide/

a. **For the inputs**: 1) *Integrity*: whether the user[20] is asking to perform a task that violates specific policies or is otherwise sensitive (e.g., illegal or regulated activity, hateful or harmful content, etc.); 2) *Bias and Stereotyping*: the possibility that users ask the model to reproduce stereotypes.
b. **For the outputs**: 1) *Misinformation*: including false, inaccurate, or misleading information; 2) *Toxicity*: content that may be perceived as offensive or insensitive, and 3) *Bias or Stereotyping:* when prompted, the model displays these traits.

However, as they make these evaluations, actors along the AI value chain lack an agreed-upon set of risks to consider. To this end, we believe NIST should provide specificity and clarity around the taxonomy of risks and considerations applicable to generative AI. This will help developers particularly as they take the preliminary steps of “defining policies”[21] and then researching relevant solutions for their specific use case.

In this context, NIST should ensure any companion guide on generative AI to the AI RMF differentiates between risks that are introduced at the level of the foundation model – for instance, a large language model, trained on a massive amount of general data, that can be adapted to a specific use case and to perform certain functions – and risks that occur when a model is integrated into an end-user product, or when given particular instructions by an end user. Differentiating in this way enables users of the guide to understand which actor in the ecosystem is best positioned to address a specific risk, and what steps are most appropriate for that actor to take. The companion to the AI RMF should also offer solutions on how to quantify and measure risks at different layers of the generative AI stack, with consideration to generative AI specific-use cases such as image generation, code generation, and multimodality. For example, for models with multilingual capabilities, additional guidance around fairness and bias in this particular context can help a company prioritize testing for particular sub-categories.

Importantly, NIST should ensure that its focus is on actionable, imminent challenges. Concerns about potential existential risks – such as the risk that AI may eventually act independently of human control, or contrary to human interests – are certainly important considerations of a broader risk management system. However, focusing exclusively on them would be short-sighted and disproportionate to identified risks, and would shift resources and attention from the imminent challenges that need to be measured, addressed, and mitigated. Additionally, a disproportionate focus on extreme risks can have significant costs for innovation and U.S. leadership, limit the benefits that generative AI could bring, and undermine the public’s trust in AI.

[20] Note: “user” in the generative AI context does not purely refer to the interaction between a person and an AI system. It can also refer, for example, to an interaction between two automated systems.
[21] Page 9 of the Responsible Use Guide, https://ai.meta.com/static-resource/responsible-use-guide/

ii. NIST should drive alignment on metrics for evaluating models

In addition to understanding what risks to test for, developers and deployers of AI models also need to have clarity around how to evaluate models against such risks and, specifically, with which metrics and benchmarks to evaluate them.

Driving consensus around the benchmarks for model evaluations should be a priority for NIST. To be effective, it is important that such metrics effectively and comprehensively capture the *impact* of the models – understanding issues in the various contexts in which they can be deployed – and not just convey a summary view of their general patterns. This will ensure a risk-based approach; certain requirements should apply *exclusively* to models that present this type of capability. In fact, a scenario where requirements are widely applied irrespective of risk – such as a scenario where pre-testing is required *ex ante* for all models – would create an insurmountable barrier to entry for smaller companies, curtail innovation, but not meaningfully mitigate actual risks. Importantly, for these models, the relevant areas of risk listed above should be incorporated into the Map and Measure stages of the AI RMF.

Additionally, NIST should focus on providing systematic and structured ways to articulate "red lines" – that is, outcomes for which to test and prevent, rather than being required to assess all conceivable scenarios – which would be impossible.

In this respect, there should be a clear distinction between models that are developed for commercial use – thus made available for use to the general public – and models that are intended for purely research purposes. In fact, releasing models to researchers, academics, and similar stakeholders is a practice that should be encouraged and facilitated, as conducive to risk mitigation and safer models. These contributors can help more quickly find and mitigate risks in systems, improve them, and fine-tuning them to prevent erroneous outputs before commercial release, which in turn makes the latter safer.

C. NIST should not overlook the value of open sourcing models for safety and security

As NIST pursues the Executive Order's core values of promoting innovation, safety, security, competition, and access throughout the lifecycle of generative AI, it should encourage, protect, and facilitate open source development and research as critical enablers of these objectives.

Cybersecurity experts and U.S. Government guidance both underscore the importance of openness and transparency to improve outcomes in security.[22] Notably, the White House Commitments include the establishment of systems for the discovery of weaknesses and

[22] Some examples are: https://www.nist.gov/publications/advanced-encryption-standard-aes-0; https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf , https://www.schneier.com/crypto-gram/archives/1998/1015.html#cipherdesign, https://blog.google/outreach-initiatives/public-policy/transparency-in-the-shadowy-world-of-cyberattacks/

vulnerabilities by third parties as an essential aspect in advancing the trust and safety of most advanced AI. In fact, by democratizing access to AI models, toxicity, bias, bugs, and vulnerabilities can be identified by an open community, and mitigations developed, to iteratively improve the models.

This is especially true in areas of security such as cryptography where it is generally understood that systems cannot be trusted *unless* there is open access to the implementation or underlying algorithm. Indeed, it is standard practice to develop cryptographic standards and implementations in the open in order to build trust in their efficacy. Additionally, open source creates a bigger, collective security team to combat bad actors who would leverage technology for negative outcomes. In fact, it lets outsiders – including academics and researchers with extensive experience – hold companies accountable which, in turn, strengthens public trust.

The benefits of open-sourcing have been proven in the cybersecurity industry. Only two decades ago, the industry shared and published significantly less information out of fear that technologies such as cryptography, technical details of software vulnerabilities, penetration testing tools, and other cyber offensive tooling would be used adversarially by bad actors. However, as these technologies and data were made more widely available through open source, good actors adopted and applied them to improve defenses.

An open and available model allows experts to test for sensitive issues within secure environments and allows them to responsibly disclose results they believe should be addressed. An additional potential benefit is that, by studying an open sourced model in-depth, experts can develop new attack vectors on other models, thus strengthening the security ecosystem overall.[23]We encourage NIST to work with other agencies to validate that an open source approach to generative AI models will allow experts to iterate faster and facilitate testing.

To be sure, Meta supports open source and its benefits for safety and security. However, we are not dogmatic about it: For this reason, we believe there's a space for a mix of proprietary and closed models on the one hand, and open and widely shared ones on the other. We come to the decision of whether to open-source or not for each model independently, and only after rigorous testing and risk assessment.

D. NIST should provide detailed guidance on red-teaming, as one method to address security and safety issues

As further discussed below, Meta places great importance on red-teaming, a type of adversarial testing which is appropriate for assessing some risks. However, we recognize that while red-teaming is useful for examining the safety of a generative AI and novel risk exploration, it is not a "one-size fits all solution." Rather, it represents one tactic that should complement other AI safety and evaluation work.

[23] https://www.cmu.edu/news/stories/archives/2023/july/researchers-discover-new-vulnerability-in-large-language-models

Furthermore, across industry there are open questions about best practices surrounding red-teaming for AI models and systems, and it is important that any red-teaming guidance take into account varying contexts in which red-teaming might occur. For example, some red-teaming exercises may be designed to test just a foundation model, whereas others test an entire AI system that includes not only the model but also user interfaces, infrastructure, and all the other elements required to deploy it in a real-world application.

In light of this, Meta recommends that NIST provide guidelines on the following topics:

- **The difference between red-teaming and other AI safety evaluation work:** Meta views AI red-teaming as one of many safety evaluation practices. This means that red-teaming is not the sole means for safety evaluation, benchmarking, or other evaluations. It would be valuable for NIST to set out guidance regarding the independence of a red-team from a safety or development team closer to an AI model in order to ensure the independence and robustness of the adversarial testing. Furthermore, NIST should look to other forms of testing as a means to address components of the relevant risk taxonomy – specifically, it should look at how different methods of measurement assess different risks, with particular deference to issues of scale and likelihood of occurrence. For example, red-teaming is well suited to identify vulnerability to bad actors and misuse, but it is also important to use other methods to assess the risk of unintended behavior in response to good-faith usage and also to ensure that risk mitigations also minimize harm to people's innocent uses of the technology. We encourage NIST to help clarify and make distinctions between different forms of measurement and mitigations.
- **Red-teaming and the AI development lifecycle:** We further recommend that NIST set out guidance on when and where in the AI development lifecycle red-teaming should take place. We encourage NIST to continue to engage in fact-finding across AI companies to identify and share best practices surrounding whitebox and blackbox red-teaming, approaches for end-to-end testing, and using safety evaluations alongside red-teaming, with the goal of proffering a holistic framework.
- **Feedback loops for novel risk discovery:** Red-teaming is valuable for conducting "novel risk research" to identify and measure risks not set out in companies' existing policies. We recommend that NIST provide guidelines on organizational feedback loops that can facilitate a company updating its risk management processes to account for new risks discovered by red-teaming.
- **Use cases, units of granularity, or deployments where red-teaming should be required in addition to safety evaluations:** This includes threat models for dangerous capabilities, guidance on when models should be red-teamed based on their deployments (i.e., public releases versus product integrations), and recommendations on when external (i.e., outside of a company) red-teaming should be considered. In this regard, it is important to distinguish between research releases and commercial releases, which carry different risk profiles.

- **Red-teaming resource investments for high-risk/high-capability models:** Because some models can produce a wide range of outputs and be adapted to a variety of use cases, more red-teaming efforts and resources (including a wider range of attacks exploring a range of outcomes) may be required to assess the risk of higher-capability models. We recommend that NIST produce guidelines on this point.
- **Automating red-teaming:** To encourage the use of technology to scale AI safety evaluation and mitigation, we encourage NIST to issue guidelines on how automated red-teaming can be used to complement manual red-teaming.
- **Metrics:** NIST should provide guidance on how organizations can determine the appropriate amount of red-teaming for a particular application. While it is intuitive that different types of models and different uses carry different risks, NIST could significantly improve the robustness and consistency of red-teaming by providing clearer metrics for determining the sufficiency of a red-teaming exercise.
- **Dangerous capabilities & CBRNE:** We encourage NIST and other USG stakeholders to work with companies on red-teaming for these risk areas, especially in light of the limited number of experts on CBRNE risk accessible to private sector entities and the legal risks surrounding testing for classified information. We would value lines of communication and expert pipelines to assist with testing.
- **Harmful content:** For red-teaming exercises that involve the production of harmful content or possibly classified information, we would value guidance on how to perform red-teaming in ways that are legal and also sufficiently robust.
- **Case studies, illustrations, and/or examples of AI red-teaming that exemplify a best practice or state of the art:** These examples would ideally encompass the lifecycle of a given AI system, for example, AI threat modeling, AI risk discovery, attacker emulation, and post-release or product launch threat intelligence.
- **The continued need for private-public partnerships on evaluations and metrics:** Meta reiterates the needs that, for certain risks (i.e., CBRNE), the government needs to work together with industry for creating benchmarks and evaluations that can be used in safety evaluations and red-teaming to identify and measure the risks posed by particular models.
- **General framework for red-teaming quality:** We would recommend the creation of a framework to assess the quality of a particular red-teaming exercise with suggestions for levers to increase the quality.

Finally, red-teaming guidelines should recognize that this practice carries a **significant cost**, which will be particularly burdensome for smaller actors. This highlights the importance of using it as a calibrated tool, useful for specific areas.

We welcome NIST's feedback on the topics, above, and any additional clarity on best practices for red-teaming.

E. Meta takes a thoughtful approach to safety and security

Meta has been building mitigations for safety and security into our products and models for more than a decade as part of our commitment to responsible AI. As we launched our latest foundation model, **Llama 2**, our approach implemented lessons learned over many years, ensuring that safeguards were built into the models from the beginning, and publicly sharing our approach in our *Llama 2 Research Paper.*[24] In particular we:

- Analyzed for bias and adopted privacy protections in the pre-training data;
- Evaluated the model against industry safety benchmarks (e.g., truthfulness and toxicity);
- Deployed red-teaming methodologies, and in particular:
    - We conducted a series of red-teaming with various groups of internal employees, contract workers, and external vendors. These teams included experts in cybersecurity, election fraud, misinformation, legal, policy, civil rights, ethics, software engineering, machine learning, responsible AI, and creative writing.
    - The red-team probed our models across a wide range of risk categories (such as criminal planning, human trafficking, controlled substances, sexually explicit content, unqualified health or financial advice, and privacy violations), utilizing a range of attack vectors (such as hypothetical questions, malformed/misspelled inputs, and extended dialogues).
    - Red-teaming results were used for further model safety training and fine-tuning.
- Provided specific guidance in our Responsible Use Guide on fine-tuning the model and red-teaming it.
- Put in place an *Acceptable Use Policy* that prohibits certain use cases to help ensure the models are used fairly and responsibly.[25]

Following the release of Llama 2, we:

- Provided downstream developers and stakeholders with reporting mechanisms for sharing violations of our *Acceptable Use Policy*.[26]
- Incentivized researchers to stress-test and report vulnerabilities though our bug bounty programs.[27]
- Submitted Llama-2 to DEFCON, a large cybersecurity conference, where 2500 hackers stress-tested the model.

Furthermore, we have implemented additional mitigations specifically for the generative AI features that Meta offers on our platforms (e.g, AI characters). These include:

[24] https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/
[25] https://ai.meta.com/llama/use-policy
[26] *See generally* github.com/facebookresearch/llama (reporting issues with the model); developers.facebook.com/llama_output_feedback (reporting risky content generated by the model); LlamaUseReport@meta.com (reporting violations of the Acceptable Use Policy or unlicensed uses of Llama).
[27] facebook.com/whitehat/info

- **Input filters**: When someone interacts with our AI features, classifiers detect whether they're soliciting content that could violate our policies.
- **Output filters**: Classifiers detect potential violations of our policies. Content that is confirmed to violate is regenerated until no violation is detected.
- **Social systems**: Content created by Meta's generative AI features, when shared onto existing social surfaces (e.g., Facebook Feed) remains subject to our broader policies, user reporting mechanisms, and existing sets of content classifiers.
- **Lockouts:** People who repeatedly attempt to generate content that violates our policies will lose the ability to interact with our generative AI features for a fixed period of time.

This experience with large language models encouraged us to share some of our expertise with the AI research and engineering community. As a result, we launched **Purple Llama**,[28] an umbrella project which includes our open trust and safety tools and evaluations. It includes both offensive (red-team) and defensive (blue-team) strategies. Purple Llama is meant to level the playing field for developers to responsibly deploy generative AI models and experiences in accordance with best practices shared in our Responsible Use Guide.

As part of this project, we released CyberSecEval and Llama Guard. **CyberSecEval** is a comprehensive benchmark developed to bolster the cybersecurity of large language models employed as coding assistants. One of the things CyberSecEval can indicate is the rate at which the model responds helpfully for requests for assistance with a cyber attack. We believe CyberSecEval is the most extensive cybersecurity safety benchmark to date. **Llama Guard,** on the other hand, is a safety classifier for filtering input and output which is trained to detect problematic or policy-violating content.

As part of our commitment to security research, we recently released **Code Lama 70B**.[29] In this model, we use CyberSecEval to understand and mitigate cyber risks in our Instruct model, prior to release, providing good-faith actors with one of the safest coding copilot tools available. This is a major step towards enabling community collaboration and standardizing the development and usage of trust and safety tools for generative AI development. We encourage other companies to release and open source their safety tools and evaluations so that others in the community can interrogate those solutions together. In addition, if NIST is well-positioned to develop these evaluations with the industry, we hope that NIST can continue to contribute to setting the standards for these technologies.

[28] https://llama.meta.com/purple-llama/?utm_source=ai.meta.com&utm_medium=redirect#why-purple-llama

[29] https://ai.meta.com/research/publications/code-llama-open-foundation-models-for-code/

III. Reducing Risks Associated with Synthetic Content

A. Industry is aligned on the need for transparency mechanisms, but not on what they should be

As technologies evolve, so can the nature and scope of the risks connected. For example, one of the main issues commonly associated with synthetic content, **mis- and dis-information,** is not a new phenomenon. However, the nature of how it is disseminated has evolved. Generative AI has made creation of deep-fakes cheaper and faster, and it is important to take proper account of this development. At the same time, generative AI has not significantly changed what resourceful bad actors can do because, even if more harmful context exists, bad actors are still constrained in their ability to distribute that content in misleading ways. From this standpoint, ongoing development of technology and policies supporting generative AI, including work on long-standing issues of detection, labeling, and policy enforcement, will be an important part of ensuring that risks associated with synthetic content are addressed effectively.

Industry is aligned on the need to build mechanisms for increased transparency, and in particular, building common technical approaches for **provenance** (e.g., watermarking). However, there is a lack of consensus on the best approach. It should be noted that all current possible technical approaches to provenance present face limitations on **robustness**. That is, they will not be sufficient to stop sophisticated actors using their own technology to circumvent transparency within AI-generated content (i.e., currently, there is no foolproof method to entirely prevent the removal of provenance measures).

For this reason, approaches that focus on harmful content without distinguishing whether that content is AI-generated or not, and are thus *not* AI-specific, continue to be crucial. For Meta, these approaches include the use of AI as an invaluable tool in detection, governance review processes (e.g., compliance with our Community Standards and Terms of Service), our fact-checking program, and educational initiatives to support people in navigating the modern information environment.

Against this background, NIST can have a pivotal role in driving consensus towards mechanisms and standards for transparency of AI-generated content.

B. NIST should leverage existing frameworks

As NIST considers how to best address the issues surrounding synthetic content, it should leverage the existing guidance and the work that has already been carried out industry-wide.

i. White House Commitments

The White House Commitments provide some guidelines with respect to issues around synthetic content. In addition to the importance of disclosure of photorealistic content, they also call on companies as responsible for developing tools and APIs able to determine whether a certain piece of content was created on their platform. Companies agreed to work together, within industry and standard-setting bodies, to develop mechanisms and systems to enable understanding of whether audio and visual content is AI-generated. This is an area where NIST could provide guidance and leadership towards consensus.

ii. PAI's Responsible Practices for Synthetic Media

Another important initiative that NIST should consider is the *Responsible Practices for Synthetic Media*,[30] driven by Partnership on AI. The Framework presents guidelines for various actors in the AI value chain: from developers, to creators, to publishers, and distributors. It encourages the development of transparency tools and the disclosure of content generated by AI, both **directly** (e.g., changing pixels to add a visible watermark to an image, identifying it as AI generated*)* and **indirectly** (e.g., changing pixels to make a machine-readable, invisible watermark, or attaching metadata to indicate that the content is AI generated, or provide other information about provenance). Meta has been collaborating with various stakeholders on this initiative, which represents an important step in ensuring transparency and guardrails around AI-generated content.

C. NIST should focus on clarifying and distributing responsibilities across the AI value chain

Safety is a shared responsibility and all actors across the AI value chain have a stake in it, including developers, deployers, individuals, civil society, and policy-makers. Mitigations should be placed at the point where they will be most effective, and upon the actor that is best positioned to implement them.

For this reason, it is not only disproportionate, but even technically infeasible to expect upstream model developers to be able to monitor every use of their model and prevent all possible risks that may arise from the synthetic content the model generates. In the same way, it would be disproportionate to expect downstream deployers to be responsible for mitigations related to the development of the AI system.

An example of a mechanism Meta has implemented, which is aligned with this idea, is our policy requiring **self-disclosure** for AI-generated content in ads related to social issues, elections, and politics. Advertisers are required to disclose whether their ads contain AI-generated imagery or video (i.e., if photorealistic or realistic-sounding audio). This functionality is very

[30] https://syntheticmedia.partnershiponai.org/

valuable to address instances in which content provenance cannot be maintained with technical measures. In this way, advertisers have an important role in transparency, alongside Meta who provides them with the tools and the relevant policies.

Determining the right obligations proportionate to the risks being addressed, and at which point those will be most effective, is an important aspect of AI governance.[31] NIST should 1) carefully assess which actor is best positioned to implement certain measures, and 2) create mechanisms to clarify responsibility in the AI value chain. The AI RMF currently includes a basic identification of the relevant actors, but does not clearly specify respective responsibilities. Moreover, in developing guidelines on synthetic content, NIST should refrain from focusing exclusively on the development of generative AI systems, and think more holistically about the responsibility of other actors to reduce risk.

D. NIST should take a differentiated approach based on the content

There is no “one-size fits all” approach from a content-type perspective. To ensure a risk-based approach is maintained, NIST should specifically focus on content that is sufficiently realistic to mislead the user, and differentiate between this content and other AI-generated materials (ex., cartoons) that are obviously fictional. This is aligned with both industry practice as well as the White House Commitments, which explicitly exclude content that is “readily distinguishable from reality” from its scope.

Consistent with this approach, Meta recommends a multi-tiered approach to watermarking photorealistic content, including development of:

- Visible markers on images;
- Markers within the metadata;
- Increased transparency and traceability, which are resilient to common image manipulations (such as cropping, color change, etc.).

These measures are included in the content from our image generator built into our Meta AI assistant, as well as appropriate in-product measures for other generative AI features. For example, in chats with AIs at Meta, people are able to access additional information about the AI, including how it generates content, the limitations of the AI, and how the data they have shared with the AI is used via in-product education. We are also developing techniques to include information within image files created by Meta AI, and we intend to expand this to other experiences as the technology evolves.

[31] https://openloop.org/wp-content/uploads/2022/10/Artificial_Intelligence_Act_A_Policy_Prototyping_Experiment_Taxonomy_AI_Actors.pdf

E. Meta is committed to transparency for synthetic content

Meta supports transparency of AI-generated audiovisual content and agrees that it is imperative that people not be misled about the source of such content. For many years, we have been thinking about similar issues, and have been working together with others to identify solutions. In 2019, we partnered with industry experts and academics to develop the *Deepfake Detection Challenge*,[32] which focused on accelerating the development of new methods to detect deepfake videos, and concluded a possible way forward to improve detection models.

At Meta, transparency for content that is generated by AI on our platforms is a priority. Amongst the solutions we've implemented in our products, the *imagine with Meta AI* experience, our text-to-image generation capability, includes visible watermarking to clearly identify content as AI generated. Soon it will include invisible watermarking applied with a deep learning model for increased transparency and traceability. While imperceptible to the human eye, the invisible watermark can be detected with automated means, particularly with a corresponding model that has the capability of reading it. The invisible watermarking is resilient to common image manipulations, such as cropping, color change (brightness, contrast, etc.), and screenshots. We aim to bring invisible watermarking to many of our products with AI-generated images in the future.

We're also investing in research and development of watermarking tools for other modalities, and on additional watermarking techniques. With respect to the latter, for example, our research activities have helped us develop *Stable Signature* – a new invisible watermarking technique to identify when an image is created by a specific generative model.[33] Stable Signature is robust to adversarial model attacks, which mitigates the potential for watermark removal. This can help identify misleading or fake images generated by open source models. The research has some limitations – for example it only focuses on images, and is only for diffusion models. It is a proof-of-concept rather than a mature solution for in-product or open source transparency. However, we are confident this effort will contribute to the development of further standards.

[32] https://ai.facebook.com/datasets/dfdc/
[33] https://arxiv.org/abs/2303.15435

IV. Advancing Responsible Global Technical Standards for AI Development

A. NIST should drive consensus and support U.S. leadership in existing standard-building processes

i. Regulatory fragmentation is a risk to innovation and should be prevented

The significant breakthroughs achieved in generative AI technology have brought it to the forefront of regulatory attention and led to multiple policymaking processes running in parallel. At an international level, these efforts include, to name a few, the EU AI Act,[34] the G7 Leaders' *Hiroshima AI Process,*[35] the *Bletchley Park Declaration*[36] promoted by the UK, the Organization for Economic Cooperation and Development (OECD),[37] the Council of Europe (CoE),[38] the World Economic Forum's (WEF)[39] ongoing work on AI governance, and the United Nations efforts led by their recently-established AI High-Level Advisory Body.[40] Domestically, the White House Voluntary Commitments and the Executive Order are only two examples of regulatory initiatives, alongside multiple legislative proposals surfacing at both a Federal and State level, and Senator Schumer's leadership in the AI Forums.

Meta shares the goal of building AI responsibly and safely. We understand that, at this juncture, there are many pending questions on how best to do so and a collective interest in getting this right. However, with so many initiatives underway, there is an increasingly significant risk that the regulatory landscape may turn into a fragmented and untenable patchwork of conflicting rules, and make it excessively burdensome, if not impossible, for companies to prioritize the key work that needs to be done to build responsibly. Without consensus on which harms and risks to prioritize, how to measure them, and work collaboratively on mitigations, this will be a disadvantage to people and enterprises. This lack may be particularly difficult as requirements layer up, as actors may play various roles in the AI value chain.

As we have seen in other aspects of global data regulation, fragmentation and an inability for governance frameworks to interoperate will have significant consequences for access to advances in AI technology, innovation, and competition, depriving societies, in particular in the Global South, of the myriad of benefits AI could bring, and potentially exacerbating existing

[34] https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai
[35] https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/g7-leaders-statement-on-the-hiroshima-ai-process/
[36] https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
[37] https://www.oecd.org/digital/artificial-intelligence/
[38] https://www.coe.int/en/web/artificial-intelligence/work-in-progress#01EN
[39] https://initiatives.weforum.org/ai-governance-alliance/home
[40] https://www.un.org/techenvoy/ai-advisory-body

issues.

We believe harmonization and common global standards are imperative, while the introduction of prescriptive domestic rules on U.S. innovation is premature, and likely counterproductive. As these ongoing initiatives mature and consensus emerges on key issues, it is important that they are given the space to succeed and to ensure consistency and interoperability across them.

ii. NIST should leverage its experience to support standards development and consensus

NIST should drive consensus and support U.S. leadership in existing standard-building processes and international organizations, as well as drive alignment across the multiple agencies entrusted by the Executive Order. We believe that, overall, the U.S. Government has taken a considerate and balanced approach in the Executive Order and in the White House Commitments, and we especially applaud the intent to further internationalize the latter through frameworks and common approaches.

The U.S. has a tremendous opportunity to lead towards this consensus, and is already doing so in multiple international fora, such as the Organisation for Economic Co-operation and Development (OECD), the G7 and the EU-US TTC. NIST can help by leveraging its expertise and drive adoption of global standards that reflect the U.S. leadership, as well as the best practices developed by industry. NIST has done this before: its *Cybersecurity Framework*, for instance, has been an example of successful standardization and global consensus, with wide international adoption across various industries.[41]

NIST's role is especially relevant in supporting the OSTP's endeavors in this area, thanks to its extensive experience in fostering the adoption of global standards, including its representation of the U.S. in the International Standards Organization (ISO). In this respect, we welcome the work of the organizations that have already mapped the NIST AI RMF with the ISO risk management standards[42] and are looking forward to continued mapping of the AI RMF with ISO 42001, so that organizations have a clear picture of how to use the framework to help execute on the standards. NIST's priority should thus be to continue supporting and fostering these standards development and alignment processes.

B. NIST should map the AI RMF to the Executive Order to promote a harmonized taxonomy

As NIST considers how to address generative AI specifically, a preliminary goal should include the alignment of definitions, taxonomies, and categorizations across the various instruments that are being developed in response to the Executive Order. Some agencies have already

[41] https://www.nist.gov/cyberframework/success-stories
[42] For example: https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

started incorporating definitions differing from those of the Order in the pursuit of their tasks, which will certainly generate misalignment and inconsistency.[43]

Given this context, one important area where NIST can have an impact is supporting the other Agencies that are provided with important tasks under the Executive Order with a common language of reference. This could be done, for instance, by mapping the AI RMF, where relevant, with the various sections of the Executive Order, especially when the language differs, but the concepts are similar.

C. NIST should leverage open source to drive inclusivity in pursuing its internationalization goals

The Executive Order tasks various agencies with the goals related to promotion of international cooperation, standardization, and information sharing. As NIST carries out part of these activities, a principle that should underpin its approach is the importance of maintaining inclusivity. Open source is essential to the U.S. goals of strengthening human rights, promoting competition, and fostering innovation. In fact, open source provides and bears the cost of building the models that smaller players can build on and benefit from. It democratizes access to pivotal technology and reduces inequality.

For these reasons, NIST should ensure it includes mechanisms that encourage and enable open innovation, as well as information sharing across companies and governments, for example, by encouraging companies to share information via research papers and system/model cards.

V. Conclusion

Meta appreciates NIST's leadership in developing consensus-based technical and risk management approaches for generative AI, and we are excited to collaborate with NIST on this important endeavor. We are confident that NIST will have a crucial impact in driving consensus toward global standards in responsible AI, fostering innovation, and promoting U.S. leadership internationally.

[43] E.g. OMB adopted a different definition of AI system than the one contained in the Executive Order https://ai.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-Public-Comment.pdf