Comments of Amazon

Request for Information re: 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) [NIST-2023-0009]

Amazon welcomes the opportunity to provide feedback on this Request for Information (RFI) related to the National Institute for Standards and Technology's ("NIST") Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (the "AI Executive Order"). The AI Executive Order seeks to establish a policy foundation for the development of AI that is "safe, secure, and trustworthy," and assigns NIST with a number of crucial tasks to advance this mission. We provide below feedback to support NIST's efforts to: (1) update the AI Risk Management Framework to account for new considerations implicated by generative AI; (2) establish guidelines to help organizations assess the capabilities of AI models and conduct red teaming; (3) prepare a report on standards, tools, and best practices for managing the risks associated with AI-generated content; and (4) develop a roadmap for global engagement to promote AI technical standards.

We are living in a unique moment of AI innovation. Advanced foundation models have given rise to large language models (LLMs) and multi-modal models that can perform a wide range of "generative" tasks across multiple domains. Generative AI is able to not only summarize and compose sophisticated text, but also generate vivid images based on textual prompts and transform a natural language description of software functionality into functional code. At Amazon we provide a broad range of capabilities and services that help customers both large and small use generative AI in creative ways, building new applications and improving how they work. For example, Amazon Bedrock enables AWS customers to build and scale generative AI-based applications by providing access to a select set of foundation models from AWS and other leading third-party model providers. Within Amazon, we are using generative AI to innovate for our customers, including through the recent integration of a new, custom-built large language model that will make the experience of interacting with Alexa far more conversational and intuitive. We are likewise providing generative AI capabilities to make it easier for our selling partners to write engaging, effective product listings, and help shoppers find what they are looking for.

Across our company, we are committed to developing and deploying AI responsibly and our response to this RFI is guided by that commitment. Last year, we proudly endorsed the White House Voluntary AI Commitments, and as described in more detail below we have taken a number of important steps to operationalize them. The Commitments set out ambitious and concrete objectives for managing many of the unique risks of powerful generative AI systems and for building trust with the public, including through red teaming, mechanisms such as watermarking to disclose AI-generated content, and research prioritization. The AI Executive Order seeks to build on these commitments, and we appreciate the opportunity to provide feedback as NIST develops guidance that will help developers and deployers operationalize them. Given the global regulatory interest in these issues, NIST's guidance can play an important role in shaping an international regulatory environment that is risk-based, technically informed, and conducive to continued innovation.

1. Updating the NIST Risk Management Framework to Account for Generative AI

The AI Executive Order tasks NIST with developing a companion resource to NIST's AI Risk Management Framework, NIST AI 100-1 ("AI RMF"), to account for the unique challenges posed by generative AI.[1] To that end, the RFI requests feedback about how organizations are currently using the AI RMF, the unique risks posed by generative AI, and the approaches organizations are taking to mitigate them. As NIST develops resources to enhance the use of the AI RMF for generative AI, it will be critical to ensure they reflect the need for a lifecycle-based approach to risk management. While responsibility for managing the risks of AI have always been shared by model providers and deployers of AI systems, the broad range of capabilities and use cases enabled by foundation models heightens the need for risk evaluations that are tied to specific end-uses.

NIST's publication of the AI RMF in 2023 was a landmark moment in U.S. AI policy. In less than a year, the NIST AI RMF has become a valuable resource for developers and deployers of AI. By establishing a shared conceptual framework for identifying and mitigating risks throughout the AI system lifecycle, the AI RMF provides a common reference point for stakeholders across the AI value chain to communicate risk. The AI RMF is particularly useful for organizations like AWS that provide tools and services that help other enterprises build their own AI applications. For that reason, AWS is committed to aligning our internal AI governance processes, across our portfolio of managed AI services, to the NIST AI RMF. We are also actively developing tools and resources to help our customers do the same, including a recent overview on performing risk evaluations that are aligned to the AI RMF and other recent standards.

While the fundamentals of the AI RMF remain sound, we agree that there is an opportunity for NIST to further increase adoption by providing companion resources to explain how the framework can be used to manage the unique risks of generative AI. In developing such resources, it will be important for NIST to convey that the lifecycle-based approach to risk management is particularly important in the context of the foundation models that power generative AI. The size and general-purpose nature of foundation models differentiate them from traditional ML models, which typically perform individual tasks, like sentiment analysis and forecasting trends. In contrast, foundation models are capable of performing a wide variety of disparate tasks, with a single model potentially being capable of natural language processing, question answering, and image classification. Notably, the highly adaptable nature of foundation models – including the possibility that they can be fine-tuned to make them more useful for yet-to-be-imagined tasks – makes it virtually impossible to anticipate the full range of deployment scenarios and use cases, which ultimately map to the level of risk involved. As a result, NIST's guidance should reflect the fact that deployers of generative AI models have a particularly important role to play in managing the risks associated with their specific use cases.

Companion AI RMF guidance should likewise provide organizations with tools and resources to help them identify, measure, and mitigate the new and unique risks that are implicated by generative AI. As Amazon Scholar Michael Kearns has noted, recent advances in generative AI have rendered some traditional responsible AI challenges more complex and given rise to entirely new technical challenges. For example,

[1] Sec. 4.1(i)(a).

generative AI systems may need additional safeguards to mitigate risks like "hallucination" (i.e., producing outputs that are plausible but verifiably incorrect) "toxicity" (i.e., generating content that is offensive, disturbing, or otherwise inappropriate), and the potential for new forms of misuse, including amplification of disinformation and exacerbating cybersecurity risks.

To address these new risks, NIST's companion AI RMF guidance should encourage organizations to evaluate the adequacy of the governance-based and technical safeguards they use to manage the risks of generative AI. At Amazon we continuously evaluate our approach to AI governance to ensure it remains fit-for-purpose as the technology evolves. For example, Amazon recently refined the internal principles that guide responsible AI development efforts to more explicitly account for the unique risks presented by generative AI. While our testing and evaluations processes have long accounted for fairness, explainability, robustness, governance, privacy and security, and transparency, we have broadened our focus in connection with our generative AI development efforts. This expansion includes three critical areas – controllability, safety, and veracity – that are key to effectively managing the risk for the foundation models powering generative AI systems. Our updated approach to responsible AI is reflected in the service card we published for Amazon Titan Text, our family of proprietary LLMs primarily designed for enterprise use cases. The Amazon Titan Text Service Card is the first of our service cards that reflects our expanded approach, providing key information about how the LLMs have been evaluated for veracity (i.e., the likelihood of hallucination), safety (including our efforts to red team the model), and controllability.

We also provide our customers with a range of resources and tools to help them manage risks and responsibilities associated with their deployment of generative AI and the underlying foundation models. For example, in November 2023, we previewed two new services Amazon customers can use to evaluate the performance of foundation models against responsible AI benchmarks and implement safeguards that are aligned to their responsible AI policies. First, Amazon Bedrock Model Evaluation offers customers model evaluation tools, including for human-led evaluations and automatic model evaluations which may be run on either their own data or public benchmarks for metrics such as accuracy, robustness, and toxicity. Second, with Bedrock Guardrails, customers are able to implement safeguards (e.g., content filters and denied topics) across their foundation models to help deliver more consistent and safe user experiences aligned with their company policies and principles.

2. Model Evaluation and Red Teaming Guidance

The AI Executive Order charges NIST with "establish[ing] appropriate guidelines. . ., including appropriate procedures and processes, to enable developers of AI, especially of dual-use foundation models,[2] to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems."[3] Red

[2] The AI Executive Order defines "dual-use foundation models" as "any model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point operations, or using primarily biological sequence data and using a quantity of computing power greater than 10^{23} integer or floating-point operations." Sec. 3(k). There is recognition that a computing power-based threshold for defining highly capable models is subject to shortcomings (*e.g.*, it could deteriorate over time as model architectures become more efficient and bad actors might seek to evade it by distributing their training workloads). However, in the absence of measurable and objective methods for evaluating model "capabilities," compute is considered a reasonable proxy measurement.

[3] Sec. 4.1(ii).

teaming is defined in the AI Executive Order as "a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI." Examples of red teaming provided include identifying "harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."[4]

Consistent with our endorsement of the White House Voluntary AI Commitments, Amazon agrees that organizations should engage in robust red teaming prior to making a dual-use foundation model generally available to the public. Red teaming such models has emerged as an important best practice because exhaustively measuring the full range of their potentially dangerous capabilities, risks, and vulnerabilities using conventional benchmarks and technical evaluation tools remains an unsolved challenge. However, the processes and techniques for red teaming dual-use foundation models are in their infancy, and evolving quickly.

A critical responsibility of NIST, together with AI developers, will be to standardize the approach for red teaming the most capable models. The aim should be to drive consensus on an evolving, versioned standard that can be quickly updated as we learn more about additional emergent capabilities of AI models and the best practices to mitigate them. As NIST contemplates potential guidance for red teaming dual-use foundation models, a critical first step will be defining the purpose and scope of a red teaming exercise. NIST should clarify that red teaming encompasses a variety of structured techniques, including the use of automated assessments, to evaluate whether a model is capable of producing dangerous or otherwise unintended outputs. The goal of red teaming should be to use a mix of automated and manual model evaluation processes to identify common patterns and vulnerabilities that may give rise to unintended system outputs. In developing guidance, NIST should recognize that automation will be key to scaling red teaming efforts and making them more measurable and reliable. Furthermore, automation may provide significant benefits, such as limiting human red teamer exposure to toxic content and allowing for an exponentially higher amount of testing to be conducted prior to model release.

NIST should encourage organizations to evaluate and red team dual-use foundation models in four key phases:

- *First*, organizations should define the scope of capabilities and risks that they are seeking to test and evaluate. Organizations should evaluate their most highly capable systems across key dimensions of responsible AI, including safety, veracity, controllability, fairness, explainability, robustness, privacy and security, and transparency. For each of these dimensions, organizations should identify a set of objective standards and technical criteria that can be used to evaluate model performance. NIST's AI Safety Institute can aid in this effort by driving consensus about how emergent risks (e.g., biosecurity) should be measured.
- *Second*, organizations should begin by evaluating their models using available benchmarks to assess model capabilities and identify potential risks. For instance, at Amazon we have used benchmarks to evaluate a model's alignment with "veracity" by measuring the model's proclivity for producing factually incorrect responses.

[4] Sec. 3(d).

- *Third*, for model capabilities and risks that cannot be evaluated using traditional testing and evaluation approaches, organizations should perform at-scale red teaming using a mix of internal and/or external resources as appropriate to manually prompt the model, with the goal of getting it to produce outputs that violate one of the standards or technical criteria.
- *Fourth*, organizations should use subject matter expert-driven red teaming to evaluate their models for risks that may require more specialized domain knowledge, such as biosecurity risks.

Importantly, NIST's guidance should encourage organizations to adopt a risk-based approach to red teaming, focusing resources on the most capable models rather than low-risk automated tools. NIST should likewise avoid issuing technically prescriptive guidance that could lock organizations into using practices that may be quickly rendered obsolete. Finally, NIST guidance should support the development of tools and benchmarks that can help organizations evaluate models for risks currently requiring specific expertise, while leaving room for organizations to adopt automated processes as they become available. NIST can contribute to the effectiveness of red teaming efforts by compiling shared datasets of prompts and prompt sequences that will help organizations evaluate models for dangerous capabilities. Collaboration with industry and independent researchers on the development of such benchmarks will be critical so that they can be evaluated and iterated upon to improve their robustness and enhance their effectiveness. NIST's recently established AI Safety Institute Consortium provide useful fora for understanding and discussing these emergent and dynamic best practices.

3. Reducing the Risk of Synthetic Content

The AI Executive Order seeks information about existing standards, tools, methods, and practices for mitigating the risks posed by synthetic content.[5] The potential for generative AI models to produce lifelike audio-visual content has prompted understandable concerns that the technology may be abused in ways that will fuel online disinformation, increase the dangers of fraud and deception, and proliferate the volume of toxic online content.

At the outset, it is important to note that there are several different technical approaches to disclosure of synthetic media, including visible and invisible watermarking, cryptographic metadata, and fingerprinting (including of image, audio-, and audiovisual content), among other tools. As with red teaming, we would encourage NIST to consider that the state of the art will evolve and any future guidance related to labeling of synthetic content should be appropriately flexible. We also encourage NIST's guidance on labelling to take a risk-based approach and to focus on deployment scenarios where confusion about the origin of content could give rise to a material risk of harm. Finally, we strongly support government initiatives to promote further research in this space.

As NIST seeks information about best practices in this area, we'd like to share some of what we are doing as an example of the types of safeguards NIST could recommend. First, we have committed that all images generated by Amazon's recently announced Titan Image Generator will contain an invisible watermark that can be validated by an API. Built-in watermarking is designed to help reduce the spread of deceptive content and disinformation by providing mechanisms to identify AI-generated images. By making the image

[5] Sec. 4.5.

validation API publicly available, our aim is to make it easier for organizations to quickly assess whether an image has been generated or augmented by Titan Image Generator. For instance, if a social media company sought to include a label on AI generated content, they could call our API to determine whether a user uploaded image was created using Titan Image Generator. AWS is among the first model providers to widely release built-in invisible watermarks that are integrated into the image outputs and designed to be resistant to alteration. To further mitigate the risk that Titan Image Generator might be misused for deepfakes or other deceptive content we have implemented a mix of training data and input filters to prevent the generation of images of real individuals. We are committed to continually enhancing the robustness of these safeguards and would welcome the opportunity to collaborate with NIST and the AI Safety Institute on these efforts.

Second, for our Amazon Bedrock service, which offers our Titan foundation models, we have applied filters on user inputs and model outputs to address a range of potential risks, including the potential for our models to produce harmful outputs. These filters use a combination of rules-based and machine learning classifiers that can be applied both to the prompt and the model output. The classifier algorithm processes model inputs and outputs, and assigns type of harm and level of confidence. The classification process is automated and does not involve human review of user inputs or model outputs and it allows us to:

- ***Filter Harmful Content***: These filters work by evaluating both user inputs and model responses against a set of defined policies and blocking them if those policies are violated. We can therefore detect and filter problematic and irrelevant queries and outputs by defining a set of denied topics (including related to safety concerns).
- ***Identify Patterns of Misuse***: We also use classifier metrics to identify patterns of potential violations and recurring behavior.
- ***Automate Abuse Detection.*** When customers use Bedrock, we may conduct automated abuse detection to detect harmful content (e.g., content that incites violence) in user inputs and model outputs. We reserve the right to suspend access to Bedrock in the event that a customer continues to use the service in a manner that may violate our or a third-party model provider's policies.

Third, our AWS Responsible Use of AI policy prohibits customers from using AWS's AI/ML services for (among other things): intentional disinformation or deception; to depict a person's voice or likeness without their consent or other appropriate rights, including unauthorized impersonation and non-consensual sexual imagery; or for harm or abuse of a minor, including child sexual exploitation.

4. Advancing Global Technical Standards for Responsible AI

The AI Executive Order directs NIST to coordinate with key international partners and with standards development organizations to drive the creation and implementation of AI-related consensus standards, and to encourage cooperation, coordination, and information sharing.[6] Consensus-based international technical standards will be crucial to promoting AI innovation across the world. In addition to their traditional role of facilitating interoperability at a technical level, standards increasingly play a fundamental role in promoting interoperability between regulatory systems. Standards provide a single source of ground

[6] Sec. 11(b).

truth that helps companies translate regulatory requirements into technical specifications that can be engineered into systems. Global standards therefore reduce regulatory fragmentation and facilitate interoperability as companies seek to build and deploy solutions in line with best practices. Technical standards also enhance trust in AI because they are inherently more agile than regulations and can more readily be adapted to address the fast pace of technological innovation.

While global standards on AI remain nascent, the International Organization of Standards reached an important milestone in December with the publication of ISO 42001, a management system standard that provides guidelines for managing AI systems within organizations. ISO 42001 emphasizes a commitment to responsible AI practices, encouraging organizations to adopt controls specific to their AI systems, fostering global interoperability, and setting a foundation for the development and deployment of responsible AI. We believe that ISO 42001 certification will be one important mechanism for organizations to demonstrate excellence in responsibly developing and deploying AI systems and applications. AWS is therefore pursuing ISO 42001 conformity and we look forward to working with customers to do the same.

Ensuring that US policy is both reflected in and supported by global technical standards should be a critical goal for NIST. To that end, NIST should ensure that any guidance it develops pursuant to the Executive Order on AI is informed by relevant technical standards that currently exist. Where technical standards do not currently exist (e.g., red teaming), NIST's ultimate aim should be to develop guidance that can be adapted into global technical standards in the future. To that end, we encourage NIST to establish formal mechanisms to ensure the U.S. AI Safety Institute is actively coordinating with international peers, including those that have established their own AI Safety Institutes.

* * *

Thank you again for this opportunity to provide feedback on the key elements of NIST's assignments under the AI Executive Order. We look forward to working with NIST to achieve these goals.