600 14th St. NW, Suite 300
Washington, D.C. 20005

February 2, 2024

Department of Commerce
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Subject: RFI Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Dear Director Locascio:

On behalf of International Business Machines Corporation (IBM), we welcome the opportunity to respond to the National Institute of Standards and Technology's (NIST) request for information (RFI) regarding the agency's assignments under Sections 4.1, 4.5 and 11 of the Biden Administration's Executive Order (EO) on Artificial Intelligence (AI). As the agency looks to implement these sections of the AI EO, we recommend that NIST:

- Make specific reference to industry best practice and mechanisms, such as FactSheets, as it develops benchmarks for mitigating AI risks;
- Prioritize transparency in all aspects of the AI Safety Institute's work;
- Leverage the AI Risk Management Framework (AI RMF) as a vehicle for promoting best practices to address synthetic content; and
- Continue engaging in regional and international governance processes to promote cross-border harmonization of AI regulatory governance, leveraging the AI RMF as a model for other countries to follow.

IBM commends NIST for seeking broad stakeholder input on its duties under the recent AI EO. We follow long-held principles of trust and transparency that make clear the role of AI is to augment, not replace, human expertise and judgement. We were one of the first in our industry to establish an AI Ethics Board, whose experts work to ensure that our principles and commitments are upheld in our global business engagements.

As you proceed, we welcome the opportunity to continue to engage to promote a responsible approach to AI that protects consumers without stifling innovation.

Respectfully,

Christina Montgomery
Chief Privacy and Trust Officer
IBM Corporation

IBM Response to OSTP RFI on National Priorities for AI

Section 1: Developing Guidelines, Standards, and Best Practices for AI Safety and Security

IBM directs NIST attention to three critical areas and activities on these topics.

First, as NIST continues to look at developing benchmarks and guidance for how best to help mitigate risks associated with AI development and deployment, **IBM urges the agency to make specific reference to industry best practices and mechanisms, including FactSheets**.[1] As an example, IBM recently released an AI Risk Atlas to assist our clients and partners in better understanding some of the risks associated with various AI models – including generative AI, foundation models, and machine learning models.[2] Coupled with other approaches, like FactSheets[3] and internal governance frameworks, efforts like this can help companies stay abreast of rapid developments in the AI threat landscape while promoting effective information-sharing between developers and deployers.

Second, **NIST should collect and promote professional skill development resources related to the development, use, and governance of advanced AI systems.** While industry best practices and governance frameworks will be a key feature to mitigate AI risks, such mechanisms are only effective if organizations have employees appropriately skilled to execute them. This profession involves some evaluation and red-teaming skills, but also things like prompt engineering (especially engineering of prompt templates and system prompts) and application development. The ethics of large-scale models in particular requires further research. The Notre Dame-IBM Tech Ethics lab has an open call for proposals on these topics and intends to advance applied research on societal, governance, and other interdisciplinary areas associated with these types of models, including the future of work and upskilling.[4]

Third, **best practices for red teaming should emphasize the importance of diverse perspectives and focus on identifying potential harms or undesirable outcomes from AI models in specific contexts and use cases**. To meaningfully evaluate whether AI models can pose risks in different contexts, red team members should come from varying socio-cultural and lived experiences. NIST should promote such diversity, as well as existing best practices for red teaming, such as leveraging comprehensive evaluation frameworks like FM-eval throughout the model's development lifecycle.[5]

[1] https://dataandtrustalliance.org/our-initiatives/data-provenance-standards
[2] https://dataplatform.cloud.ibm.com/docs/content/wsj/ai-risk-atlas/ai-risk-atlas.html?context=wx&audience=wdp
[3] https://aifs360.res.ibm.com/
[4] More information on the call for proposals can be found here: https://techethicslab.nd.edu/call-for-proposals/
[5] https://www.ibm.com/downloads/cas/X9W4O6BM

Overall, as NIST works to develop guidelines for best practices and standards for AI safety and security, it should leverage and build on its past successful work in this space. IBM supports the development of a companion resource to the NIST AI RMF for generative AI and values NIST's commitment to gathering input from relevant stakeholders. IBM has provided more specific details on practices for organizational implementation of AI RMF core functions related to generative AI through its participation in the NIST Generative AI Public Working Group (GAI PWG).

Section 2: Reducing the Risk of Synthetic Content

Individuals have a right to know when they are directly interacting with an AI system, and organizations should prioritize policies that create that awareness for consumers.[6] IBM has long supported mechanisms for promoting transparency in the use and deployment of AI. Additionally, AI developers can benefit from similar transparency in code development practices. That is why IBM offers the capability for clients to enable automatic tagging of AI-generated code in our watsonx Code Assistant product.

And although mechanisms for distinguishing between human- and AI-generated content are notoriously difficult to develop and implement, at IBM we believe that technology solutions show progress in this domain and warrant increased investment. For example, IBM helped developed a framework called RADAR, which leverages AI to detect AI-generated text that can typically otherwise evade detection.[7] Solutions to the problems of synthetic content will continue to be a flashpoint for policymakers and technical experts alike, which is **why we recommend NIST continue leveraging the AI RMF as a vehicle for promoting best practices in this domain and prioritize work on this topic in the US AI Safety Institute.**

Section 3: Advance Responsible Global Technical Standards for AI Development

AI Risk Management and Governance

IBM was pleased to be involved in the development of the NIST AI RMF, offering contributions based on our experience as a global leader in implementing AI risk management and governance processes.[8] We continue to support this related work and strongly encourage NIST to promote the use of the framework. It can help companies operationalize an AI governance system based on risk, leveraging the collective experience of industry.

For example, our own experience building and operationalizing these systems is why so many other companies trust IBM's experience and expertise in the AI governance arena. To make informed decisions about the AI tools they adopt, to monitor whether AI performance is trustworthy, to avoid costly missteps, and to take advantage of the productivity gains AI provides, IBM helps companies employ a three-step approach to AI governance:

[6] https://newsroom.ibm.com/Whitepaper-A-Policymakers-Guide-to-Foundation-Models
[7] https://radar.vizhub.ai/
[8] https://www.ibm.com/impact/ai-ethics

1. **Build the foundation for AI oversight.** Organizations need to begin their AI journey by assessing their existing strengths and weaknesses in integrating AI tools and managing the risks associated with the technology's use. This also necessitates having a clear set of goals and priorities and involving multidisciplinary teams to assist in crafting the broader AI governance strategy.

2. **Document your ethics.** Every organization needs a set of documented and widely-available principles that help to inform the design, development, and/or deployment of AI tools internally. At IBM, our Principles for Trust and Transparency are further supplemented by our *Ethics by Design Playbook*, which helps teams and individuals across the company put our principles into practice.[9] Additionally, organizations should consider how and when they can best make use of tools, such as algorithmic impact assessments for high-risk applications of AI, to help guide them in developing and deploying this technology.

3. **Adapt existing governance structures for AI.** Effective governance of AI does not necessarily require new institutional structures to support implementation. Many organizations can leverage existing governance programs (e.g., third-party risk management, procurement, legal and compliance, etc.) to effectively manage risk. At IBM, we have operationalized our AI governance via a "hub-and-spoke" model, with our Office of Privacy and Responsible Technology serving as the "hub" and the various stakeholders in our business units and regions serving as the "spokes." As AI technology, policy, and governance are evolving arenas, the key to any successful AI governance structure is to build flexibility and adaptability into the process.

US Leadership on AI Safety Standards

Establishing US leadership on AI safety standards will be critical to promoting the safe development and deployment of AI both domestically and globally. NIST's AI Risk Management Framework is considered a gold standard by other countries pursuing their own frameworks for addressing AI risk due to its comprehensive, inclusive, and transparent development process that solicited feedback from a broad and diverse stakeholder community.

IBM appreciates the coordination among government agencies around AI risk management and governance, including State Department's promotion of the AI RMF globally. We are supportive of the general approach taken by the G7 Hiroshima Process, as well as other emerging frameworks, such as the draft ASEAN "AI Guide" and the recently completed NIST-IMDA "crosswalk" between Singapore's Model AI Governance Framework and NIST's AI Risk Management Framework. **The United States should continue engaging regional and international processes to promote cross-border harmonization of AI regulatory governance with continued emphasis on the NIST AI Risk Management Framework as a**

[9] https://www.ibm.com/downloads/cas/7PWAWRQN

model for other countries to follow. Additionally, IBM offers the following three recommendations for NIST to consider as it continues to undertake this important and timely work.

First, over the past year, significant headway has been made in developing greater international consensus around AI definitions and a common understanding of the provisions that should be included in AI governance frameworks. For example, many organizations and countries have made use of the OECD's AI definition as the foundation of their AI governance frameworks. Given that definition's increasingly widespread adoption, we suggest NIST also promote its use. Another example is the work on international standards taking place in ISO/IEC JTC 1 under the secretariat of the United States national standards body, ANSI. A number of standards are already available or under development (for example, on AI management systems, concepts and terminology, risk assessment and governance.)

Second, the work of the US AI Safety Institute (AISI) will be more scientifically and technically involved than the work of the AI RMF, but should similarly prioritize transparency and inclusivity. This will lead to deepening our understanding of AI safety risks and effective mechanisms and processes for mitigating those risks, as well as ensure broader adoption of the resulting AI safety standards. The creation of the AISI Consortium is a welcome development and sends a strong signal that NIST's work on AI safety standards will follow the successful model of the AI RMF. However, as the work of the AISI is broader than just the Consortium, **NIST should prioritize transparency in all aspects of the AISI's work and clearly communicate methodologies, priorities, findings, and other work products developed beyond the Consortium.** NIST should also collaborate with the newly formed UK AI Safety Institute, European standardization organizations, and other bodies focused on this topic to ensure an open exchange of information that can help to promote more harmonized approaches to AI standards development.

Third, NIST should continue to coordinate with other agencies, including the AI Safety Advisory Board that the Department of Homeland Security (DHS) has been tasked to establish, just as CISA has been working with other nations' secure by design frameworks inclusive of AI.[10] These existing global linkages might be leverageable here to at a minimum avoid duplication of work or conflicting requirements.

Conclusion

IBM commends NIST for its longstanding work on this topic. We thank you for considering these comments and welcome the opportunity to engage with the agency as it moves forward in this process.

[10] https://www.cisa.gov/securebydesign