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

September 7, 2021

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

Subject: "Artificial Intelligence Risk Management Framework," Docket #210726-0151

Dear Acting Director Olthoff:

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 "Artificial Intelligence Risk Management Framework" (hereafter, "AI RMF").

IBM has long been committed to responsibly developing and deploying new technologies, and artificial intelligence (AI) is no different. The proliferation and adoption of new innovations require trust – between companies and their clients, consumers and technology providers, and citizens and their government. This NIST process is an important step in helping to foster a more robust and trustworthy ecosystem of AI development and adoption. We believe this consultation is an important step in promoting the principles of trust and responsibility among a wide range of stakeholders, as well as showcasing the United States' leadership as a paragon of the values that underpin trustworthy AI.

We thank you for your efforts in commencing the process of developing an AI RMF and look forward to both your consideration of these comments and future opportunities to contribute to the work ahead.

Respectfully,

Christina Montgomery
Vice President and Chief Privacy Officer
Co-Chair, IBM AI Ethics Board
IBM Corporation

Francesca Rossi
IBM Fellow and AI Ethics Global Leader
Co-Chair, IBM AI Ethics Board
IBM Research

IBM Response to NIST RFI – "Artificial Intelligence Risk Management Framework"

IBM welcomes NIST's efforts to solicit stakeholder input for the development of the AI RMF. In general, we agree with NIST's choice of the 8 attribution categories, which will serve as a solid foundation upon which to develop the specific elements of the AI RMF.

At this stage, we believe NIST's efforts would be best served by focusing on developing a coherent methodological approach for better defining how actors can assess risk along multiple inflection points in the AI development pipeline. In other words, IBM would recommend that NIST focus on determining how the AI RMF should consider practical questions, such as assessing risk, and later build on that foundation by tying it to more technical questions like explainability and transparency. This consultation effort – and future workshops and engagements – will be an important foundation for better understanding the broader industry experience in the development of AI.

In service of contributing to this early-stage approach, IBM will focus the following comments on responding to the 12 questions solicited by the AI RMF RFI, drawing from our experience in the AI field.

1. The greatest challenges in improving how AI actors manage AI-related risks—where "manage" means identify, assess, prioritize, respond to, or communicate those risks.

Controlling and managing risk in the pipeline of AI development is a highly distributed challenge, primarily due to four factors: (1) the roles of different actors vary greatly across different engagements and business models;[1] (2) the technology sector is highly dynamic, with large variations in business models that change over relatively short periods of time relative to other sectors; (3) AI is a highly dynamic field and model development processes, as well as applications, change over time based on context and environmental exposure; and (4) AI is a versatile technology, which means many of the most significant risks are use-case dependent.

Indeed, some risks materialize much further downstream the development pipeline and may not be widely recognized by every actor, which makes it difficult for AI developers to manage, and therefore mitigate, those risks in the early stage of the technology's lifecycle development.

[1] We would draw distinctions between those entities that design and develop AI systems (developers/providers), those that adopt and deploy AI systems (deployers/owners), and those groups or individuals who are responsible for the oversight and ongoing monitoring of an AI system (end-users). For more details, see Ryan Hagemann, "Precision Regulation for Artificial Intelligence," IBM Policy Lab, 21 Jan. 2020, available at https://www.ibm.com/policy/ai-precision-regulation/; see also "Confronting Bias: BSA's Framework to Build Trust in AI," available at https://ai.bsa.org/wp-content/uploads/2021/06/2021bsaaibias.pdf.

This is especially true when considering the fluidity and diversity of roles and responsibilities in the AI ecosystem. The purpose of an AI system may be dictated by a consumer-facing corporate user or the developer from which it purchases the system. A system may be developed entirely by one party, or it might encompass AI tools purchased from multiple vendors and patched together by a systems integrator or the user's employees. Training data may be obtained by a user internally or from vendors, and a system's sensitivity to new training data over time, or lack thereof, might dictate that the original developer or a later user should test for bias and accuracy. Many of IBM's own AI products are general-purpose tools and APIs that clients use to develop and train, with their own data, AI systems across various industry sectors.

The dynamism of this marketplace is certain to persist into the future as stakeholders experiment with different technical approaches and business models to maximize the benefits of AI for society – all of which will need to be accounted for when considering how best to identify risks and assign responsibilities accordingly.

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI.

IBM has long been committed to the need for trustworthy AI, which we define as a human-centric approach to developing and deploying AI that rests on three core ethical principles: (1) AI should be used to augment human intelligence, (2) data and their insights belong to their creators and owners, and (3) AI deployed in the marketplace should be both transparent and explainable, relative to the user's expectations and the use-case application.[2]

As the AI RMF process moves forward, it will be important to ensure all stakeholders are working from the same shared language. To that end, we would suggest the following definitions for "explainability," "transparency," "robustness," and "fairness."

- **Explainability.** AI explainability is the ability of an AI system to provide a human-interpretable explanation for its predictions and produce insights about the causes of its decisions.

- **Transparency.** AI transparency refers to the ability of AI systems to share information on how it has been designed and developed. Examples of this information is what data is collected, how it will be used and stored, and who has access to it; test results for accuracy, robustness and bias; and the kind of explainability which is provided by the system. Transparency does not entail

[2] See https://www.ibm.com/watson/trustworthy-ai.

companies revealing source code or other forms of trade secrets or IP. Instead it focuses on making the purpose and the properties of an AI system clear to users.

- Fairness. Fairness in connection with an AI system refers to the equitable treatment of individuals or groups of individuals. Defining equity depends on the context in which an AI system is used. For example, lack of diversity in training data may lead to biased output in the context of an AI system making recommendations that impact opportunities available to human beings).

- Robustness. AI robustness is the property of an AI system that allows it to handle exceptional conditions, such as abnormalities in input, effectively. AI robustness also allows AI systems to respond well to deliberate adversarial attacks, minimizing security risks and enabling confidence in system outcomes.

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability.

As a starting point, we believe these three principles are the bedrock upon which an effective AI RMF should be built. We would suggest "robustness" as another principle worthy of consideration. If a system is unacceptably vulnerable to attack or otherwise unable to produce replicable output within its intended field of use, then transparency, fairness, explainability, and accountability are likely to be at risk.

4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management—including, but not limited to, the management of risks related to cybersecurity, privacy, and safety.

At IBM, we take our commitment to trustworthy AI seriously. Although we recognize that technologies are tools that can be used for good or bad, it is also true that companies have a responsibility to help ensure their innovations are used for the betterment of society. In order to meet that value obligation, IBM stood up an internal AI Ethics Board to hold ourselves accountable to the promises we make to society. The IBM AI Ethics Board is a centralized body composed of a cross-disciplinary team of professionals that aims to foster a culture of ethical, responsible, and trustworthy technology development and implementation. Its remit is to provide a centralized governance mechanism for reviewing and issuing decisions regarding our technology ethics policies, practices, communications, research, products, and services.

We are also incorporating a Privacy and Security by Design (SPbD) approach to the creation and development of our products and services, from internal tools to Cloud and Software-as-a-Service offerings. The goal of SPbD is to create a standardized set of best practices to which our technologies are subject, including architecture reviews and security testing, to ensure we are well-positioned to meet the ever-changing demands of the regulatory landscape. We build adherence to SPbD practices by disseminating guidance to help

product teams understand the expectations of the SPbD discipline and group those documents into related topics to provide internal stakeholders with a foundational context before providing more detailed descriptions of specific sub-disciplines.

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above.

IBM has an array of tools that can provide a methodological foundation that NIST can consider leveraging as it moves forward in developing the substantive elements of the AI RMF. We would point to our AI Factsheets 360 toolkit. We define a FactSheet as "a collection of relevant information (facts) about the creation and deployment of an AI model or service." That information could be things like the purpose of an AI model, measurable characteristics pertaining to the underlying dataset, the model itself, or steps taken by developers during the creation and deployment of the model.[3] IBM has also contributed to the development of numerous other toolkits, such as AI Fairness 360 and AI Explainability 360, that can also be used to help identify, assess, and mitigate AI-associated risks.[4]

Although the methodology we describe is about how to determine what information to include in a FactSheet, the process has broader applicability to NIST's AI RMF efforts – in particular, the process is geared towards getting organizations to consider what information is of primary importance for a given model and what should be measured to assess any associated risk. For more information on the specific steps involved in this methodology, we recommend NIST review our attached paper, *A Methodology for Creating AI FactSheets*.[5]

We would also urge NIST to strongly consider the elements incorporated into BSA | The Software Alliance's *Framework to Build Trust in AI*.[6] This first-of-its kind risk management framework is the product of a multi-industry effort to help inform practices for mitigating unintended bias in AI systems' lifecycle developments and creates a methodological framework for conducting impact assessments that help to identify and manage risks.

6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles.

IBM strongly believes that AI is a tool that can help supplement human decision-making, not a replacement for human judgement. There are many existing regulatory regimes that include reporting requirements intended to address and safeguard social priorities,

---

[3] See https://aifs360.mybluemix.net/introduction.
[4] For more details and information on other toolkits, see https://developer.ibm.com/blogs/ibm-and-lfai-move-forward-on-trustworthy-and-responsible-ai/.
[5] Also available at https://arxiv.org/pdf/2006.13796.pdf.
[6] Available at https://ai.bsa.org/wp-content/uploads/2021/06/2021bsaaibias.pdf.

including such disparate examples as public financial disclosures aimed at protecting investors and product safety information for equipment that poses possible risks to users.

Reporting practices for AI should be focused on providing required information for users to optimize the value of an AI's output in order to perform the roles for which they remain accountable. This could include, among other things: ensuring people know when they are interacting with an AI system, disclosing the diversity of a training dataset's demography, or detailing context- and user-specific explanations of an AI output or its limitations.

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts.

As NIST moves forward to develop the AI RMF, it will be imperative for the agency to factor existing standards-setting processes into its framework development process. In particular, we would draw NIST's attention to the forthcoming ISO AI Management System Standard (ISO/IEC JTC 1/SC 42/WG 1), which is currently in development and targeted for a January 2022 publication date. This standard aims to define a general procedure for when and how an organization should consider the use of AI risk and/or impact assessment.

There are many other sector- and industry-specific rules, regulations, and soft law (e.g., informal guidance, standards) that may govern the development and use of AI in specific contexts. The Supervisory Guidance on Model Risk Management (SR-11-7), for example, aims to "provide comprehensive guidance for banks on effective model risk management" for "a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates."[7] While not specific to AI systems, guidance such as SR-11-7 is an example of an effective soft law approach to governing risk using an outcomes-based, technology-neutral set of guardrails and best practices.

As part of NIST's AI RMF consultation, we recommend that future RFIs, workshops, and engagements make identifying and categorizing these types of soft law systems a priority, as they can help inform the broader effort at producing a viable AI RMF.

8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.

As noted previously, IBM's AI Ethics Board takes a lead role in addressing and managing the implications of our technology's use. The AI Ethics Board serves as an indispensable component of our governance framework for internal review of products and offerings that

[7] See https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107a1.pdf.

require additional risk and value assessments, helping to manage potential safety and ethical concerns regarding particular use-cases.

IBM has also been vocal in our support of testing for bias in AI. Most recently, we outlined a specific set of obligations that companies developing high-risk AI systems should be adhering to, including:

- Requiring bias testing and bias mitigation, in a robust and transparent manner, for certain high-risk AI systems such as law enforcement use cases. These high-risk AI systems should also be continually monitored and re-tested;

- Focusing any requirements for conducting impact assessment prior to deployment on owners of those high-risk AI systems that pose the greatest potential to harm;

- Documenting the assessment processes in detail, making them auditable, and retaining them for a minimum period
of time;

- Convening and driving national and international forums to accelerate consensus around clear and consistent standards, definitions, benchmarks, frameworks, and best practices for trustworthy AI;

- Providing resources and expertise to help all organizations—not just large corporations—ensure their AI is deployed responsibly;

- Increasing investment in research and development around bias testing and mitigation to ensure leading scientific approaches are targeted at mitigating bias; and

- Supporting accelerated developer training around bias to ensure appropriate training aimed at understanding and mitigating biases and recognizing how bias could be unintentionally introduced into AI systems during the development pipeline.[8]

9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, "AI RMF Development and Attributes").

IBM believes the 8 attributes NIST has proposed are both appropriate and sufficient as a bedrock upon which the AI RMF's more technical details can be built.

10. Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational

[8] Dr. Stacy Hobson and Anjelica Dortch, "Mitigating Bias in Artificial Intelligence," IBM Policy Lab, 26 May 2021, available at https://www.ibm.com/policy/mitigating-ai-bias/.

processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include—but are not limited to—the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation.

As discussed previously, we would recommend NIST review our attached paper, *A Methodology for Creating AI FactSheets*, and the associated AI FactSheets 360 methodology linked to above.

More generally, we would draw NIST's attention to the numerous existent national governance frameworks, such as Singapore's Model AI Governance Framework and the ongoing work at the OECD, that prioritize risk-based and outcomes-oriented approaches based on consensus-based multistakeholder engagements. In addition, we would warn against the instantiation of artificial categories (e.g., "users") in favor of real-world functions (e.g., providers of training data, defining use-cases) to reflect the need for specificity. Given that such functions tend to vary substantially by individual use-cases and business models, it will be important for NIST to focus on the need for specificity and clarity in order to develop a clear pathway for practical implementation of the elements that will come to make up the AI RMF.

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

The AI RMF can provide a meaningful opportunity to emphasize the need for human-centricity in not only the development of AI, but in how it is utilized by "humans-in-the-loop" post-deployment. For the accountable individual in such a role to put the appropriate weight on an AI's recommendation(s), it is important for the human in question to understand the AI's capabilities and limitations. For any given dimension of risk, there are numerous examples of how a more diverse, knowledgeable, and skilled workforce can help mitigate that risk. The AI RMF process should include a discussion, and accompanying assessment, of industry experiences in the workforce development space – the challenges, opportunities, and lessons learned.

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Any AI RMF will need to address broader issues related to governance – not speaking to formal regulatory oversight and enforcement, but rather looking to best practices and recommendations. We believe this will be a necessary conversation as part of the AI RMF development process, as in the absence of good governance, even a robust framework may

suffer over time due to undetected programmatic weaknesses. To appropriately scope this conversation in future consultations, however, we would recommend that “governance” be narrowed to describe the way those measures are undertaken to create fairness, transparency, and accountability are tested, monitored, and defended.