Bipartisan Policy Center Response to NIST on Artificial Intelligence (AI) Risk Management Framework Draft

The Bipartisan Policy Center is committed to developing viable, consensus-driven solutions to improve AI standards and ethical frameworks and appreciates NIST's invitation to inform the Artificial Intelligence Risk Management Framework (AI RMF) development. BPC works with a wide range of stakeholders from government, academia, industry, and civil society to develop recommendations for AI that are responsible, equitable, and accountable. We are pleased to share our expertise and research in the comments below for consideration in the Framework.

We appreciate NIST's undertaking of such an important assessment framework and work organizing a consensus-driven, open, transparent, and collaborative process. After reviewing the initial draft AI RMF and participating in the workshop discussions, we would like to submit these written comments for consideration when developing future drafts.

Our comments will address four primary areas of concern regarding the draft AI RMF:

1. Identifying and defining common terminology for the design, development, use, and evaluation of AI systems through the voluntary framework;
2. Specifying laws, regulations, standards, guidelines, or best practices that interact with this voluntary framework and reinforce the mapping, measuring, and governance practices;
3. Establishing robust processes for measuring and defining impact and introducing feedback mechanisms to assess risk; and
4. Conveying an agile process for stakeholder engagement for ongoing use of the framework over the lifecycle of an AI system.

Detailed Response

1. ***Identifying and defining common terminology for the design, development, use, and evaluation of AI systems through the voluntary framework***

Takeaways and recommendations:

- Common taxonomy will facilitate discussions about AI risks
- Add a distinct definition of "impact" and its relation to "risk"
- Clearly define "harm" and present specific examples

Throughout the second workshop on Building the NIST AI Risk Management Framework, a recurring concern was a need for common language and definitions of key terms used in the framework. The initial draft of the framework builds a taxonomy around AI risk using technical characteristics, socio-technical characteristics, and guiding principles. This initial draft creates a comprehensive approach to assessing AI's characteristics but does not provide a detailed terminology to facilitate discussions between different organizations about the associated risks

and how to mitigate harm. One panelist explained that a complete taxonomy is necessary to define AI risks, facilitate work between organizations, and communicate risks to consumers.

The initial draft of the framework has contradictory meanings of "risk" that may cause uncertainty and weaken the framework. Though generally, risks have negative associations, the framework "intends to offer approaches to minimize anticipated negative impacts of AI systems and identify opportunities to maximize positive impacts," building off the ISO's definition that maintains certain risks can be positive. If risks yield positive and negative results, then the later characterization that "trustworthiness and risk are inversely related" is incorrect. NIST must clarify the definition of "risk" and separately and distinctly define "impact" and its relation to risk.

There are other considerations when defining AI risks. Risks can evolve and be weighed differently by one person in an organization than another. Additionally, internal decision-makers may weigh risks differently than external stakeholders who experience those risks. The intended use of the AI system may also produce risks that differ from risks assessed during the initial design. The framework should clearly define the resulting "harms" or "potential harms" from an AI system to help an organization evaluate and govern its AI.

2. ***Specifying laws, regulations, standards, guidelines, or best practices that interact with this voluntary framework and reinforce the mapping, measuring, and governance practices***

Takeaways and recommendations:

- Identify best practices, standards, regulations, laws to include in an AI Resource Center
- Provide examples of how different industries can use the framework

Clear communication about the framework's objective and how it applies in tandem with other policies is detrimental to its success. An AI Resource Center, or a similar hub, will streamline access to the resources needed to maintain safe and effective AI systems.

An AI Resource Center should include the following documents:

  - ISO/IEC 22089
  - ISO/IEC 42001
  - ISO/IEC 23894 potentially utilized by any size org, customizable
  - OECD AI Classification Framework
  - IEEE Ethically Aligned Design
  - P7000 suite of standards
  - IEEE P3119 procurement of AI and automated systems
  - ISO/IEC 24030 use cases

The framework can also play a critical role in shaping the development and use of AI in sectors where the risks of AI are pertinent. Existing vertical regulations in sectors such as housing,

insurance, employment, and others may not provide the oversight necessary to protect individuals from the risks of AI. Active application of the framework and future standards and regulations will be vital to protect individuals in high-risk areas. Examples of how it might apply to different sectors may help organizations realize the value of adopting an AI risk management framework.

3. *Establishing robust processes for measuring and defining impact and introducing feedback mechanisms to assess risk*

Takeaways and recommendations:

- Better define how to measure impact and how that definition will feed into future AI policies
- Stronger language in Table 4 about feedback mechanisms from a diverse group of users, both the direct userbase and those potentially impacted
- Clarify the role and responsibility of each level to judge risks and audit systems

The framework helps characterize risks; however, NIST must standardize how impacts and risks of AI systems are measured. As drafted now, the framework cannot guide robust measurements of AI impact and risks. There are no criteria on which to evaluate the risks of an AI system or conditions in which risk is too harmful to an individual or society. The European Commission's Artificial Intelligence Act, for example, proposes a risk-based approach to strictly define AI risks and prohibit some AI systems. Some concerns with this approach are that it theoretically may deter developers of AI from innovating. Opposingly, some argue this approach will yield only safe uses of AI and these benefits will outweigh the costs. On the other hand, NIST's framework largely lacks values for different impacts or risks. While this approach gives developers and organizations greater autonomy over developing their AI systems, organizations may not effectively evaluate risks in their AI systems using only this framework. We are not endorsing any specific approach, but a close review of the benefits and weaknesses of each approach is important to draw a compromising solution.

NIST could also clarify its feedback processes in the AI framework to ensure governance practices are inclusive of all impacted stakeholders throughout the lifecycle of an AI system. The framework supports good governance practices referenced in Table 4, such as, "Organizational policies and practices are in place that prioritize the consideration and adjudication of external stakeholder feedback regarding the potential individual and societal harms posed by AI system deployment." However, the framework fails to identify feedback processes for those users not involved in developing an AI system. The framework should consider providing more details about how individuals from a diverse group of users, both the direct userbase and those potentially impacted, could communicate their concerns.

Additionally, the framework indicates that the responsibility for deploying safe and responsible AI systems falls on every member in an organization, but most significantly on the executive

leadership. However, the framework does not clearly define who is responsible for auditing the system during development and after deployment. The next iteration of the framework should determine responsibility so organizations can prepare individuals or teams to perform these practices and prepare leaders to take responsibility for implementing ongoing auditing practices.

4. ***Conveying an agile process for stakeholder engagement for ongoing use of the Framework***

Takeaways and recommendations:

- Establish and communicate a process to collect feedback and revise the framework if necessary
- Create incentive models for using this tool and incorporating it as part of the enterprise model
- Establish accountability for governance function over a lifecycle of an AI system

NIST has generated significant recognition and legitimacy for the framework by encouraging stakeholder involvement in its development and ongoing discussions, including requests for information, workshops, and other means of stakeholder engagement. These efforts will continue to improve the framework so long as NIST considers stakeholders' concerns and implements compromising solutions that will enhance the framework.

We encourage NIST to continue its efforts to incorporate feedback and promote open dialogue with a diverse and multidisciplinary set of stakeholders long after the final framework is released. Methodologies to map and measure risks are still maturing, and societal expectations are constantly evolving; therefore, NIST should be responsible for establishing a process to collect feedback and revise the framework if necessary.

The framework's longevity will depend on voluntary use over the lifecycle of AI systems. Because the framework is voluntary, corporations may need more substantial incentives to adopt NIST's framework in tandem with internally produced AI risk management frameworks, AI risk assessments, or standards and guidelines. Additionally, NIST should clarify how organizations should utilize the framework over the entire lifecycle of an AI system. We appreciate the work by NIST to produce such a fundamental framework and encourage additional efforts to promote sustained use of the framework across industries and organizations.