BIPARTISAN POLICY CENTER

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

Introduction

BPC 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 development. BPC works with a wide range of stakeholders from government, academia, industry, and civil society to develop recommendations for AI, and we are pleased to share our expertise and research in the comments below for consideration in the Framework.

The Bipartisan Policy Center played a central role in developing a national strategy on artificial intelligence for Congress, working with Reps. Robin Kelly (D-IL) and Will Hurd (R-TX). The bipartisan AI national strategy passed the House of Representatives as concurrent resolution H.Res.1250. Much of our response for this RFI reiterates critical points from a series of reports to supplement that national strategy. In these reports, we advocated for Congress to authorize and provide robust funding to NIST to develop voluntary standards frameworks to help address bias and fairness issues based on a cooperative and multi-stakeholder approach.

We look forward to NIST's continued undertaking of similar efforts to address these issues.

Goals of This Request for Information (RFI)

1. Identify and better understand common challenges in the design, development, use, and evaluation of AI systems that might be addressed through a voluntary framework;
2. Gain a greater awareness about the extent to which organizations are identifying, assessing, prioritizing, responding to, and communicating AI risk or have incorporated AI risk management standards, guidelines, and best practices, into their policies and practices; and
3. Specify high-priority gaps for which guidelines, best practices, and new or revised standards are needed and could be addressed by the AI RMF—or which would require further understanding, research, and development.

Detailed Response

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:

One of the greatest challenges is determining the level of risk associated with different applications of AI. Using a risk-based approach, as prescribed by the Office of Management and Budget, organizations can determine which AI-related risks are acceptable or have the potential to cause unacceptable harm.[1] This approach accepts that AI actors will take some inevitable risks but requires actors to be transparent about their evaluations of risks, fostering both accountability and innovation. Deep, sector-specific knowledge is necessary to evaluate solutions and understand associated risks sufficiently. This approach should also be context-specific, and the weight of an AI actor's best judgment should vary by sector and application. Highly sensitive areas such as health care, lending, criminal justice, or housing should be treated differently than low-risk areas.

A diverse workforce with a broad perspective and understanding of risks associated with AI applications is necessary to identify, prioritize, and respond to risks. The challenge of creating a diverse workforce for AI requires a holistic approach, starting from early education and throughout a career. It must focus not just on recruiting talent but also on developing and retaining existing talent, which requires looking at an organization's culture and whether it is inclusive. This includes diversifying organizations' leadership. Research has been done to understand the prevalence and consequences of discrimination in AI systems. Efforts are also underway to find ways to mitigate human bias and help under-represented communities and marginalized groups realize their potential[2] using AI. Different interventions should be tailored for optimal effect at different points in the pipeline, from education through to careers.

These systems are dynamic and can change every day. Efforts to mitigate risks should include accessing results, as well as the process to get said results. In the case of online systems perpetually updated through users' engagements, a system could be ethical and fair or unethical and discriminatory depending on the day, the user, or the context. Inherently these changes in ethical and unethical results carry enormous risks for the outcome of the product.

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,

---

[1] Draft White House OMB Memorandum on Regulation of AI (January, 2019). https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

[2] 11449 (mit.edu)

security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

Alongside the risk-based approach prescribed by the Office of Management and Budget, NIST should consider how different 'accuracy' and 'fairness' metrics can be applied to different use-cases. In the context of binary decision-making processes, for example, accuracy is increased by every correct prediction made, despite the consequences of incorrect decisions carrying different costs depending on the social context. As just one example, an AI system designed to detect cancerous growth in using MRI images could either incorrectly diagnose cancer when none exists (leading to unnecessary and costly further medical procedures) or incorrectly deem a patient healthy when cancer exists (leading to delays in treatment and adverse health outcomes). Every decision-making process makes decisions about which of these errors to prioritize[3] - NIST's AI RMF should include an analysis of the potential downstream effects of the different types of errors an AI system makes.

Similarly, NIST's AI RMF should consider establishing guidelines for metrics of fairness in different contexts. It has been shown that many metrics for fairness can be mutually exclusive, leading to situations where different stakeholders can come to wildly different conclusions when analyzing the same data[4]. NIST can provide guidelines on which fairness metrics should be employed given the context and risk level of each AI application, guiding the trade-offs that exist for each.

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;

Per NIST's commitment to foster public confidence and trust, we believe ethical principles of artificial intelligence should be considered in the framework. The Department of Defense (DOD) recently adopted five ethical principles of AI to advance trustworthy AI technologies after receiving input from experts from industry, government, academia, and the American public. They call for responsible, equitable, traceable, reliable, and governable AI for combat and non-combat purposes. These principles can be used to inform NIST's development of ethical principles in AI, and both the DOD and NIST should continue working closely with industry and experts to develop and refine guidelines for implementing the ethical principles of AI. Greater research and development of ethical principles of AI will also help the U.S. accelerate and sustain global leadership in AI.

After receiving input, NIST should ultimately balance the differences in stakeholders' definitions and approaches to ethics, with an eye towards instilling confidence in the public's

[3] https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/
[4] https://arxiv.org/abs/1811.07867

perception of AI systems. AI ethics should be centered around human ethics. A collaborative effort is necessary to determine what ethical values society wants AI to reflect. These should be centered on core principles that are defined and mapped to the ethical framework. While the risk framework is developed separately, we hope that NIST will connect this product to ethical principles developed with the input of AI stakeholders.

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;

As our world becomes more dependent on technology, greater consideration must be given to the ramifications of technology in our organizations and how it impacts the wellbeing of people and society. The previous version of NIST's Risk Management Framework was published on January 16, 2020, and we appreciate NIST's openness to continue receiving feedback on this topic.[5]

We recommend that an AI risk management framework be tailored specifically for the use of AI, and solutions should reflect the immediate risks presented by AI. Privacy, cybersecurity, and safety are all critical factors to integrate into an AI risk management framework.

AI risk management will present new considerations for organization leaders and could require additional education and training to understand how standards should be implemented. The educational system and workforce training programs should be reworked to ensure all organizations have the tools to employ this framework successfully.

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;

NIST should consider ongoing efforts in the federal government to develop and implement a Federal Data Strategy. It includes provisions for mitigating risk associated with confidential and sensitive data collected by the federal government. The effort has worked diligently to convene stakeholders from civil society, the government, and federal agencies.

[5] NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management

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;

As of April 2021, the European Union has proposed a set of regulations on using AI systems according to a risk-based classification system similar to NIST's RMF.[6] Beyond ensuring that the risk framework is interoperable with the one developed in the EU, NIST should explore how to mitigate the criticisms of the EU model.

The Federal government deploys oversight of federal IT networks through implementing the Federal Information Technology Acquisition Reform Act (FITARA), passed by Congress in December 2014. Every year agencies are graded using the FITARA Scorecard, which creates categories for measuring IT modernization efforts. Once developed, the AI Frisk Framework should use this mechanism to ensure that federal agencies apply the standard to their AI programs. NIST should work with the House Committee on Oversight and Reform, the committee that administers FITARA, every year to update guidance for agencies to implement the framework.

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;

The Consumer Technology Association has developed an AI standard for healthcare applications called the ANSI/CTA-2090. While BPC does not endorse the standard from an organizational perspective, it recognizes the comprehensive efforts by the signatory organizations to create a standard for their industry. It employs a consensus-driven standard that considers three expressions of how trust is created and maintained: human, technical, and regulatory.

Over the past several years, state governments have passed or are considering AI legislation. NIST should endeavor to capture as many of the risk definitions as possible from these state actors. NIST can be "the" risk standard makers, and state agencies should adopt federal standards due to the porous nature of data.

A robust stakeholder engagement model should be deployed to ensure that NIST captures relevant considerations for its framework. This will mitigate risks for deploying the framework.

[6] .https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

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.

AI technologies can have the potential to benefit people and do so in an inclusive manner, such as allocating resources to those most in need. But AI systems built on historically bias data or non-representative samples can have serious risks, such as an AI system picking up and exacerbating human biases, worsening inequities, and harming those who are most vulnerable. Because an imbalance of representation in the AI industry risks perpetuating historical inequalities[7], the framework should encourage organizations to increase involvement in AI development among races, genders, and marginalized people to reduce the risk of negative impacts on those underrepresented groups. Organizations that successfully develop a diverse and inclusive workforce can improve future innovative efficiency.[8]

Greater investment is needed to research and develop ways to reduce the risks of bias. One technical solution, explainable AI, or interpreting how a model reaches specific decisions can foster transparency and trust. It can also be a crucial tool for determining whether a system suffers from bias, and if so, whether technical or implementation methods can be employed to reduce or eliminate it. Other, non-technical solutions should also be used to mitigate the risks posed by AI on underrepresented groups.

The additional investment must be made to increase inclusiveness in the evaluation and testing processes of AI technologies. Leaders representing a majority population may be uninformed of an AI system's unintended biases on a minority or underrepresented group. Inclusivity built into a review process for AI technology would help mitigate this risk and promote trust. Though, traditional evaluation methods may not work to test the risks of AI technologies. For instance, the Institute for Defense Analysis has developed AI-specific characterizations to test AI systems: formal methods, cognitive instrumentation, adversarial testing, and run-time monitoring[9]. NIST should continue to explore methods that improve diversity in AI talent and mitigate risks posed to underrepresented groups.

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

[7] https://ainowinstitute.org/discriminatingsystems.pdf
[8] http://web.pdx.edu/~jizhao/MayerWarrZhao(2017).pdf
[9] https://www.ida.org/-/media/feature/publications/t/th/the-status-of-test-evaluation-verification-and-validation-of-autonomous-systems/p-9292.ashx

1. Be consensus-driven and developed and regularly updated through an open, transparent process.

We applaud NIST's work to develop voluntary standards frameworks to help address bias and fairness issues based on a cooperative and multi-stakeholder approach. We look forward to NIST's continued undertaking of similar census-driven efforts to address these issues.

2. Provide common definitions

The framework can be used to help develop common language and terminology to guide discussions about how to incorporate evolving societal values into AI design and to characterize aspects of AI risks and trustworthiness. Through our research, we found a lack of consensus on the meaning of the terms fairness, bias, or privacy, and expect that society will continuously debate how to define these terms in the context of AI. Providing one definition will prove complicated – terms such as fairness and trustworthiness can never truly be defined mathematically, and because AI systems need instructions to operate, any attempt to encode fairness into an AI system will be imperfect.

3. Use plain language that is understandable by a broad audience

In setting standards for AI terminology, public engagement is necessary to ensure diverse perspectives are considered. Mindfulness about the diverse context in which an AI system is used and its various features are also essential.

4. Be adaptable to many different organizations, AI technologies, lifecycle phases, sectors, and uses.

The framework should evaluate needs from a range of industry experts and be adaptable to new sectors and future uses of AI technology. For instance, the transportation industry will likely change dramatically with the rise of autonomous vehicles and other AI-enabled technologies. Regular input from stakeholders, such as car manufacturers, passengers, pedestrians, local officials, and academics, can help ensure a framework is adaptable given such changes.

5. Be risk-based, outcome-focused, voluntary, and non-prescriptive. The framework should focus on the value of trustworthiness and related needs, capabilities, and outcomes. It should provide a catalog of outcomes and approaches to be used voluntarily, rather than a set of one-size-fits-all requirements, in order to: Foster innovation in design, development, use and evaluation of trustworthy and responsible AI systems; inform education and workforce development; and promote research on and adoption of effective solutions. The framework

should assist those designing, developing, using, and evaluating AI to better manage AI risks for their intended use cases or scenarios.

6. Be readily usable as part of any enterprise's broader risk management strategy and processes.

7. Be consistent, to the extent possible, with other approaches to managing AI risk. The framework should, when possible, take advantage of and provide greater awareness of existing standards, guidelines, best practices, methodologies, and tools for managing AI risks whether presented as frameworks or in other formats. It should be law- and regulation-agnostic to support organizations' ability to operate under applicable domestic and international legal or regulatory regimes.

8. Be a living document.

NIST should consider developments in AI technology and new risks to society and continue to update this framework based on collaborative feedback from industry, academia, government, and civil society.

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 National Security Commission on Artificial Intelligence final report states that the AI talent gap impedes the U.S. government from becoming AI-ready[10]. Based on consensus from stakeholders, BPC recommends the following five principles be added to the framework to ensure the American workforce and agencies can thrive in the AI-driven economy:

1. The United States should embrace and take a leadership role in the AI-driven economy by filling the AI talent gap and preparing the workforce for future jobs. However, in doing so, policymakers should make inclusivity and equal opportunity a priority.
2. Closing the AI talent gap requires a targeted approach to training, recruiting, and retaining skilled workers. This AI talent should ideally have a multidisciplinary skill set that includes ethics.
3. Federal agencies leverage some of the largest data repositories to implement programs that directly impact the American population. These impacts are magnified once an AI system is employed. Given this scope, efforts must be made to recruit and retain a workforce with the skillsets needed to use AI responsibly. That starts by making it easier for people with the right skills to apply and be hired at these agencies.

[10] https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

4. The AI talent gap is not the only challenge of the AI-driven economy, so the federal government should focus more broadly on future jobs and skills complemented by AI technology. Additionally, encouraging workers to develop basic AI and technological literacy can help them better determine how to complement AI systems.
5. The educational system from kindergarten through post-college is not yet designed for the AI-driven economy and should be modernized.
6. The skills in demand in the future will continuously change, so lifelong learning and ways to help displaced and mid-career workers transition into new jobs are critical for the future workforce[11].

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

We determined that good governance structures and proper regulatory frameworks for privacy are critical for building trust in AI technologies. However, governance guidelines should be regularly reviewed to better understand how organizations can better promote AI technology's ethical use and design.

U.S. government involvement in standards-setting and measurement is important because policymakers will have direct access to the quantitative information needed for better, evidence-based decision making. It further helps government experts identify areas where targeted grants—such as academic research on quantifying AI "robustness" and "trustworthiness"—would help establish well-defined and effective metrics. At the same time, the federal government should invest in research, development, testing, and standardization to build and deploy more trustworthy cutting-edge AI systems.

Conclusion

NIST will play a vital role in fostering trustworthiness in AI, a technology that will significantly shape our future. BPC's response to the AI Risk Management Framework provides recommendations developed from collaboration with industry, academia, government, and civil society on these topics and should be considered in combination with the other responses that have been submitted. We strongly recommend that the framework constantly be reviewed and modernized as the technology continues to develop. BPC looks forward to continued work with NIST to collaborate on these concepts.

[11] https://bipartisanpolicy.org/report/ai-the-workforce/