BIPARTISAN POLICY CENTER

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

Jeremy Pesner, Sabine Neschke, John Soroushian

The Bipartisan Policy Center (BPC) is committed to developing viable, consensus-driven solutions to improve AI standards and ethical frameworks and has been covering such frameworks and laws for several years now. In 2020, BPC produced a series of whitepapers to contribute to an overall national AI strategy for the United States. More recently, we have turned our attention to AI impact assessments, and hosted four different events, bringing in a variety of stakeholders to discuss the issue throughout the year. All of them stressed the need for clearer guidelines and greater accountability. We have recently published a report on the European Union's forthcoming AI Act and its possible implications for U.S. policymaking. We specifically noted that there is the potential for a "Brussels Effect," in which the process of AI developers complying with EU regulations makes them the default in the U.S. (We have seen this as a direct result of the European Union's GDPR, with many websites accessed from the U.S. asking users' cookie preferences.) We followed up with an event focused squarely on the subject, which included a representative from NIST. Throughout our experiences in AI, we have found that many stakeholders feel that clear, unambiguous guidance is lacking in this space. That is what the RMF has the potential to become.

We appreciate that NIST read many of our previous comments and incorporated them into this version of the AI Risk Management Framework. Our previous comments centered around the need for defining common terminology related to AI evaluation and use, specifying other laws and frameworks that interact with this one, introducing processes for measuring risk and impact, and considering an agile process for stakeholder engagement. To each of these points, we note that the AI RMF Resource Center "will include a knowledge base of trustworthy and responsible AI terminology" (p. 8), that "contributed guidance may address issues including but not limited to how the Framework can be used with other AI risk management guidance" (p. 3), that "approaches for measuring impacts on a population should consider that harms affect different groups and contexts differently. AI system impact assessments can help AI actors understand potential impacts or harms within specific contexts" (p. 14), and "Measurable continuous improvement activities are integrated into system updates and include regular stakeholder engagement" (p. 31).

We are pleased to see many other developments made to this draft AI RMF. NIST has made a variety of additions and changes that indicate it continues to consider these issues deeply. The development of the Govern-Map-Measure-Manage framework in its Playbook breaks down the fundamental steps and aspects of AI governance into understandable, distinct portions of AI's development lifecycle. It is effective as a more digestible version of the RMF, but as we will discuss, we believe it will be strengthened with more tangible and grounded directions. We also appreciate the acknowledgment of other AI laws and guidance, including the EU's forthcoming AI Act and the OECD Framework for the Classification of AI systems. This recognizes that many AI considerations do not stop at national borders and therefore require harmonized governance worldwide.

The inclusion of a website for the Playbook alongside the RMF allows for a more intuitive and visually appealing design and communication mechanism than a linear written report. We hope that all important aspects of the RMF will be incorporated into the Playbook's website and other materials. Other engaging means of communicating this material could include video interviews with relevant stakeholders, short educational modules like those found on sites like Coursera, and even a simulation in which users can be presented with challenging situations to try and resolve. These options can all help bridge the gap between guidance and lived experience.

We also note and encourage NIST's further elaboration of the particular kinds of risks endemic to AI systems. Identifying the categories of valid, safe, fair, secure, transparent, explainable, and privacy-enhanced helps to specifically outline the expectations that NIST and other actors hold for trustworthy AI systems. These are all essential criteria for AI developers and operators to keep in mind as their systems are deployed. We also appreciate mapping these characteristics to other frameworks, demonstrating the harmony between them and the need for AI systems to behave consistently regardless of the region of the world in which they operate.

However, while we acknowledge the benefit of these references, we also wish to see this RMF more explicitly implement, adopt, and incorporate details of existing AI impact and governance frameworks. This goes beyond their inclusion in the proposed resource center, as they should be centered within the main document itself. AI impact assessments provide a series of specific, targeted questions for AI development teams to ask themselves about their projects' structure, goals, and assessments. There are a variety of existing impact assessments whose questions and directions can be directly incorporated into the RMF. Some of them include the Canadian government's Algorithmic Impact Assessment tool, the U.S. CIO council's own Algorithmic Impact Assessment, and the European Union's Assessment List for Trustworthy Artificial Intelligence. If a government or organization may wish to create a new impact assessment, it is advised that the assessment capture risk, cover the entire lifecycle, operate in a multistakeholder fashion, and assist with go/no-go decisions. Aside from impact assessments, which tend to emphasize the risks and potential downsides of such systems, it is also helpful to employ systems that consider their benefits as well. Ben Shneiderman's Human-Centered AI emphasizes the process of developing and testing the AI and the product designed to augment rather than replace human performance. This allows us to see AI governance both in the frame of mitigating risks and expanding the use of helpful systems.

Integrating such frameworks and impact assessments into the core RMF allows it to be more specific, grounded, detailed, and easily understandable in assessing risk. While the top-level values specified by NIST, such as "safe," "fair," and "transparent," represent clear aspirational goals and are likely to be agreed upon by everyone in principle, it proves difficult in practice to ascertain when and whether they have been met. A report from Open Loop indicates that, of the AI companies they partnered with to create a policy prototype, "most focus[ed] solely on risks related to the design and operation of their system such as dataset bias and performance issues – i.e. functional risks – as opposed to a broader set of risks related to the ethical application of automated decision-making (ADM) systems, and the societal effects of these decisions such as impact on human well-being, fairness, human interaction, end user autonomy, or overreliance on AI/ADM systems – i.e. structural risks. This program further demonstrated that a procedural approach to risk assessment, where organizations identify, assess, and mitigate risks by following a series of steps, indicative criteria, and examples, can be an adaptable alternative to a prescriptive regulatory approach applied to specific business sectors or intended uses." (p. 5) In other words, a series of clear steps and questions, such as those that impact assessments provide, appears to better contribute to risk assessment and overall AI governance. Such processes often help make matters tangible and clear, especially to those who are not well-versed in estimating social impact. This RMF will serve as a guide for best practices, but best practices require a focus on the practical – *how* does NIST recommend AI companies meet these expectations?

We appreciate that the aim of the RMF is to avoid negative impacts. However, in the event that such impacts occur, responsible parties should be defined and accountable. This point was included in our previous round of comments, and we hope NIST will consider including an accountability framework in the core RMF. From data collection and organization to algorithm design to testing, there are many points at which an AI system may begin providing problematic results. Therefore, it is important that all actors in AI development have clear standards to which they can be held. While we hope that they will always adhere to these standards, they should know that there will be consequences if they do not. One potential framework for this is the World Economic Forum's Empowering AI Leadership report, which describes several "lines of defense" in AI governance among different actors involved in its development. Some may disagree with the precise ordering, but this demonstrates that all stakeholders in AI development must not only take credit for its benefits but also take responsibility for its problems. Adding a category of organizational responsibility in the "Manage" section would help to communicate its importance.

We appreciate NIST's continued attention to and development of this important framework. This is a major step forward in AI governance and management, and by publishing this RMF, NIST places itself at the forefront of a critical debate in the United States. We applaud the public outreach and work NIST has done to collect feedback from stakeholders through public comments and workshops. We hope that the next version of the RMF will set the stage for AI governance and management in public and private institutions alike.