Insight

Federal AI Action: The First Step Toward NIST Guidance on AI Development

JOSHUA LEVINE | FEBRUARY 14, 2024

Executive Summary

- In response to President Biden's Executive Order (EO) Concerning Artificial Intelligence (AI) the National Institute of Standards and Technology (NIST) held a Request for Information (RFI) on developing guidance for the development, deployment, and evaluation of AI technologies.
- NIST's RFI marks the beginning of a regulatory process that could have major implications for the way AI models and technologies are developed, deployed, and evaluated, which has prompted some AI proponents to raise concerns about the EO's potential negative impact on the production and diffusion of the technology.
- As NIST crafts such guidance, the agency should prioritize flexibility and an iterative approach to standards, draw from existing strategies and protocols being deployed by the private sector and other stakeholders, and align guidelines with international and allied-nations frameworks related to AI-powered technologies when beneficial.

Introduction

On October 30, 2023, President Biden issued an Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. One of the tasks established in the EO is for NIST to commence a Request for Information (RFI) to contribute to the development of guidelines, standards, and best practices for AI safety and security, reducing the risk of synthetic content, and advancing responsible global technical standards for AI development. These broad categories include specific questions related to economic and security implications of content provenance and watermarking (tools for identifying and tracking generated media), best practices for red-teaming (internal teams that attempt to breach or expose cybersecurity vulnerabilities in a model or platform), procedures for mitigating harm related to synthetic content, and opportunities for NIST's guidance to contribute to international consensus covering the development and deployment of AI.

The EO included directives to agencies to enact rulemaking to develop guidance and standards for the development, deployment, and evaluation of AI-powered technologies and models. The RFI is an initial step for developing guidance that could have a profound impact on the choices made by AI model developers as well as firms and government entities deploying AI-powered technologies moving forward. There has been pushback to the EO from Republican lawmakers and groups associated with the technology industry, arguing the EO centralizes power to regulate a spate of emerging technologies and could harm future investment and innovation.

In our comments for the record to NIST, Danny Doherty and I addressed questions within each section of the RFI, emphasizing the benefits incorporating existing and developing standards from industry, civil society, and multi-stakeholder groups could have on the guidance's diffusion and impact. First, guidance for the economic

and security implications of watermarking and content provenance should prioritize existing protocols and strategies being developed by multi-stakeholder organizations and incorporate the NISTs AI Risk Management Framework (RMF) and Cybersecurity Framework. Second, NIST should incorporate traditional cybersecurity risks and red-teaming activities as well as emerging techniques specifically designed for generative AI models. Third, NIST should allow ongoing litigation or legislation to take effect before providing guidance on mitigating harms from synthetic content. Finally, when drafting guidance, NIST should consider areas where standards already do or can further align with allies to limit regulatory arbitrage and promote economic benefits of standards alignment.

President Biden's EO and NIST RFI

President Biden issued the EO in October 2023 to direct federal agencies and begin government action to promote safe, secure, and trustworthy development and use of AI. Generally, the EO tasks federal agencies to study how AI can be developed and deployed to improve government functions while respecting individual's civil rights and minimizing potential harm. The EO is largely focused on the federal government's role in developing and deploying AI, but there are some actions that would impact the private sector.

Some have raised concerns that the EO would limit innovation and development of AI-powered technologies. A group of Republican attorneys general are expressing concerns that the president's EO centralizes government control over innovative technologies and improperly uses the Defense Production Act to compel firms to disclose certain information about training runs and model development. Further, developers, firms, and civil society groups have taken issue with provisions that would increase the government's ability to regulate certain forms of model development based on compute thresholds. There are also concerns that the EO draws heavily from proposed legislation, usurping congressional authority and further expanding the power of the executive branch.

The AI EO also directs agencies to gather information to inform future rulemaking and standard setting. NIST's RFI is the first step in the process and requests information on three big-picture topics: developing guidelines, standards and best practices for AI safety and security, reducing the risks of synthetic content, and strengthening American leadership abroad. In our comments, we highlight how existing standards and other soft-law mechanisms can be building blocks for addressing concerns related to the proliferation and impact of AI models, particularly generative models.

Contributions to the RFI

Economic and Security Implications of Watermarking, Provenance Tracking, and Other Tools

First, the RFI asks about the economic and security implications of tools designed to identify content as synthetic, most notably regarding content provenance and watermarking, it should start by evaluating existing practices such as collaboration between industry and civil society. Recent reports from Goldman Sachs and McKinsey illustrate the economic potential created by AI-powered technologies and generative AI, respectively. "Content provenance" is a technical practice that labels data and creates a traceable chain so users and platforms can evaluate the evolution of a piece of content over time. "Watermarking" helps identify whether an image was created or modified by an AI-agent by embedding pixels or an invisible stamp of sorts into visual or audio content. As we highlight, collaboration between industry and civil society has already led to significant developments on these fronts, and any guidance should leverage tools such as the Coalition for Content Provenance and Authenticity (C2PA), the World Wide Web Consortium (W3C), and Google Mind's SynthID.

Regarding security implications, our comments emphasize that NIST should focus on addressing two specific challenges: 1) mitigate harm created by models, and 2) address security vulnerabilities within models. For the former, concerns related to the verifiability of content can be addressed using some of the techniques noted above, as well as NIST's RMF and red-teaming to limit a model's ability to provide such outputs. For vulnerabilities within models, the NIST Cybersecurity Framework should be a central feature of any guidance. Further, many firms leading the development of foundation models such as OpenAI, Anthropic, Meta, and Google have published information on their red-teaming activities that could help inform policymakers on the types of tactics currently being employed.

Red-Teaming and Information Sharing

In addition to tools for identifying content, the RFI asks about how red-teaming can mitigate harms created by the model and address security vulnerabilities within the models. Red-teaming has traditionally referred to techniques or attempts to expose security vulnerabilities within software. In the generative AI context, much of the discussion has focused on "prompt hacking" or trying to manipulate models so they provide outputs that violate guardrails or behaviors that developers attempted to mitigate. Our comment recommends that NIST guidance place AI red-teaming into two buckets. The first, AI red-teaming, would deal with issues related to intellectual property, privacy, and security vulnerabilities of AI models. The second could be referred to as AI testing, which would focus on the safety and outputs of AI tools to ensure they are functioning properly based on prior training and stress testing. By bifurcating these evaluations, NIST could bring together traditional elements and expertise in cybersecurity with more novel approaches that accompany the rise in generative AI and AI-powered technologies.

Further, guidance should encourage information sharing between public, private, and civil society stakeholders. A useful case study could be the Financial Services Information Sharing and Analysis Center (FS-ISAC). The FS-ISAC has developed standards to protect privacy and security for the global financial system by engaging with firms and organizations involved in global financial transactions. Partnerships related to AI are already forming, such as the C2PA noted above, as well as the Partnership on AI, which brings together leading AI and technology firms with think tanks and non-profits, media organizations, and academic institutions to propose frameworks and craft best practices for securing and monitoring models. It may be beneficial to draw ideas from the work of these organizations regarding how to best promote information sharing related to security.

Reducing Risks of Synthetic Content

NIST should refrain from issuing definitive guidance or recommendations related to reducing risk of synthetic content outside of earlier recommendations for promoting watermarking and content provenance because of ongoing litigation and legislative interest. As our comment notes, there are real potential harms that could be

created using generative models and questions of where liability should lie. Specifically, there is active litigation between foundation model developers and artists, writers, and other creatives who claim that model training and certain outputs constitute copyright violations. Further, there have been recent instances of deepfake content being created and disseminated with the intent to harm or deceive others, such as deepfake pornography of high school students and deepfake audio of President Biden urging New Hampshire voters to stay home on election day. While these are serious concerns, in addition to active litigation, members of Congress have proposed several bills attempting to address these issues, and the U.S. Copyright Office has solicited comments to inform future rulemaking. Considering the existing activity, NIST should wait for further legal or legislative action to inform future guidance.

Crafting Standards to Promote Innovation and Trust Globally

As with the internet, global competition and innovation is occurring rapidly in the realm of AI-powered technologies. While many leading models are from American companies, there are European and Chinese competitors, with users across the globe taking advantage of these new offerings. Consensus standards related to business best practices such as transparency, data collection and storage, and information sharing can be important for promoting economic growth, adoption, and further innovation. As NIST works to develop standards, it should integrate information and ideas from NIST's AI RMF as well as the U.S. Governmental National Standards Strategy for Critical and Emerging Technology (NSS-CET)because these frameworks provide foundations for evaluating the tradeoffs of global standardization in the context of the development, deployment, and incorporation of AI-powered technologies.

Weighing the Tradeoffs Related to Global Consensus Standards

While harmonization of standards and regulations has its benefits, there can be considerable costs when harmonization would harm AI development. Specifically, our comment highlights how regulations put forward by the EU and the People's Republic of China (PRC) could be out of step with U.S. interests related to the development and deployment of AI-powered technologies. The EU's General Data Protection Regulation, for example, imposes negative effects on innovation and new firm formation, as well as high compliance costs for small and medium firms. The EU's looming AI act is looking to similarly create a regulatory "floor" for AI development and deployment globally. Similarly, China's attempt to bend standards to the benefit of its national champions by passing laws regulating the types of training data and acceptable outputs would clash with more light-touch, permissive frameworks being embraced by the United States and allied nations such as the UK, Japan, and Israel. This does not mean the United States should ignore opportunities to harmonize with the EU or the PRC, but any analysis should consider how proposals such as those from the EU and China could be detrimental to the development and deployment of AI-powered technologies.

Conclusion

If approached wisely, NIST can lay the foundation for standards that could help promote innovation and competition within the market for AI-powered technologies. By providing guidance for developers and deployers on issues related to model development, red-teaming and stress testing, appropriate data practices, and other issues, NIST can encourage firms and individuals to build models that are safe, secure, and trustworthy. Further, as guidance is developed, looking to the private sector, multi-stakeholder organizations, and other nations can provide useful references in areas such as watermarking and content provenance. To achieve the goal outlined by the AI EO and the RFI, NIST should consider how previous documents and standards it and other bodies have promulgated can contribute to the present and future iterations of guidance.