Center for American Progress
1333 H Street NW, Suite 100E
Washington, DC 20005
202.682.1611
americanprogress.org

February 2, 2024

Laurie E. Locascio
Director of the National Institute of Standards and Technology (NIST) and
Under Secretary of Commerce for Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Submitted electronically via www.regulations.gov

Re: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11); Docket Number: 231218–0309, December 21, 2023

The development of artificial intelligence (AI), and in particular the public availability of generative AI and the widespread ability to automatically generate images, audio, and video, is already having a significant impact on society. The development and deployment of generative AI is happening at a speed and scale that is likely to exceed previous technological deployments.[1] Unfortunately, this rapid deployment has meant the risk management and trust and safety features that would traditionally have time to develop do not currently exist.

On October 30, 2023, President Biden signed the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[2] (AI EO) which directed NIST to undertake several tasks including "developing a companion resource to the AI Risk Management Framework, NIST AI 100-1, for generative AI."[3]

As CAP has previously stated, voluntary risk management frameworks are not a sufficient substitute for needed AI regulations and legislation.[4] Scholars have noted the shortcoming of a risk management framing for AI.[5]

However, as AI legislation or regulation faces an uphill battle in the United States in the immediate future, voluntary frameworks like the NIST AI Risk Management Framework (NIST AI RMF)[6] can be a first step in helping to identify and potentially mitigate harms from Generative AI. **As NIST carries out the mission assigned to it by the AI EO, the NIST AI RMF generative AI companion and any updated NIST AI RMF should:**

- **Incorporate the White House Blueprint for an AI Bill of Rights[7] (AI Bill of Rights).**
- **Define and include requirements for the responsibilities and risk management for developers of AI models and first- and third-party deployers of those AI models.[8]**
- **Adopt the categories from the draft OMB AI guidance where AI use is presumed to be Safety-Impacting or Rights-Impacting[9] and craft risk mitigations for these categories.**
- **Prioritize recommendations to address generative AI's risks to the integrity of elections and democratic processes given the historic number of elections taking place in 2024.[10]**

Below CAP provides the following response to the Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11); Docket Number: 231218–0309, December 21, 2023. Please contact Adam Conner (aconner@americanprogress.org: 202-669-5671) with any questions.

Sincerely,

The Center for American Progress

Adam Conner
Vice President, Technology Policy, The Center for American Progress

Megan Shahi
Director, Technology Policy, The Center for American Progress

The NIST AI RMF generative AI companion and any updated NIST AI RMF should incorporate the AI Bill of Rights

In October 2022, the White House unveiled the Blueprint for an AI Bill of Rights that included five principles: (1) Safe and Effective Systems; (2) Algorithmic Discrimination Protections; (3) Data Privacy; (4) Notice and Explanation; and (5) Human Alternatives, Consideration, and Fallback.[11] These principles are a roadmap to "guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence."[12] While the NIST AI RMF[13] does not currently reference or incorporate the AI Bill of Rights, the NIST AI RMF generative AI companion and any updated NIST AI RMF should incorporate the principles of the AI Bill of Rights as key risk management measures. CAP has previously called on the administration to incorporate the AI Bill of Rights into an AI executive order and into legislation.[14]

There is an urgent need to define and include requirements for the responsibilities and risk management for developers of AI models and first- and third-party deployers of those AI models.

This RFI specifically identifies as a priority "Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users)."[15]

A key component to upholding values of responsible AI, and specifically generative AI, is the articulation of requirements and enforcement of said requirements for developers and deployers of AI models. While governments, companies, and civil society are actively considering how to build AI responsibly, in practice the primary focus has been on first-party usage. In response to the responsible AI moment, model developers have implemented some modest safety measures at the model level—and additional trust and safety features in their own first-party deployments of their AI models—while third-party deployments via APIs have limited to no safety requirements. The lack of oversight and enforcement of third-party generative AI usage poses unique and imminent risks that governments and companies must prioritize mitigating. There is an urgent need for a standardized framework to ensure responsible use and deployment of generative AI, encompassing both first-party and third-party applications. This framework should prioritize user safety, transparency in policy enforcement, and accountability for both developers and deployers.

Thus far, when discussing risks of generative AI, various reports, governments, and other bodies have utilized different terminologies to describe the roles involved in developing and deploying AI. Unfortunately, we have not found standardized definitions for terms. For our recent February 2024 CAP report "Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone"[16] (attachment 1 at the end of this comment) we developed our own working definitions to help describe and articulate the issue. The glossary below is copied that report offers definitions of key terms to unify their meanings and work toward building a shared understanding of these systems:[17]

- ***Developers:*** *Entities or individuals involved in the creation and development of AI systems. Developers are responsible for the foundational work of building, training, and refining AI systems, such as large language models (LLMs), that power generative applications. In some cases, a single entity may function as both a developer and a deployer, managing the entire process, from AI model creation to its application and user interaction.[18] For example, OpenAI is the developer of the ChatGPT LLM, and Google is the developer of the Gemini LLM.[19]*
- ***Deployers:*** *Entities or individuals that implement and manage AI technologies in user-facing applications or services. Deployers typically use the tools and models offered by developers, primarily through an application programming interface (API), to provide AI-driven services or features within their own products or platforms. This includes the integration of AI functionalities into apps and optimization of the user experience.[20] Historically, those building using APIs and on platforms are also called developers, but in this report, "developers" refers only to those companies who built the AI models.*
- ***First-party AI systems:*** *AI systems that are hosted and operated by the developer of the AI-based technology. These entities not only develop the AI models but also manage their deployment and user interaction on their own platforms, such as websites or apps. For example, Google has developed the Gemini LLM, which is used to power Google's Bard chatbot.[21]*
- ***Third-party AI systems:*** *Entities or individuals that are external and independent from the original developer of AI systems. They are deployers of the AI systems and may use the AI technology in various applications, offer analytical insights, or develop derivative services based on the original technology.[22] Often, they are accessing the AI model via an API. For example, Snap Inc. uses OpenAI's ChatGPT via API to power its My AI bot in its app Snapchat.[23]*
- ***Open-source AI models:*** *AI models whose underlying source code, design, model weights, and/or training methods are made publicly accessible via open-source licenses. Meta's Llama 2,[24] Mistral AI,[25] and BigScience's BLOOM[26] are examples of open-source large language models.*

NIST should prioritize standard definitions for these common terms to avoid confusion and publish them in the AI RMF generative AI companion, any updated NIST AI RMF, and the NIST Computer Security Resource Center glossary.[27]

Widespread use of generative AI carries risks that must be appropriately mitigated by deployers and developers. The lack of adequate mitigations to safeguard systems, be transparent with users and stakeholders, and uphold responsibility for tools can be characterized as lackluster at best and dangerous at worst. The API access component of generative AI represents the lion's share of a developer's growth, scale, and profit potential,[28] but is not matched with similar degrees of trust and safety investment by the platforms. The nature by which generative AI has rapidly scaled to hundreds of millions of users[29] means that developers will not be afforded the same multi-decade timeframe they had with legacy products, such as social media, and they therefore must prioritize building accountability, liability, and transparency into their systems right away.

Developers have built some reporting, controls, and general protections for first-party usage, but these safeguards still lack the robustness and detail to effectively mitigate risks and protect users. For example, OpenAI, Microsoft, Meta, Anthropic, and Google have acceptable use policies for their generative AI tools.[30] These usage policies include important prohibitions on the "generation of malware"[31] and the "planning or development of activities that present a risk of death or bodily harm to individuals."[32] But neither the usage policies nor additional documentation[33] enumerates in any detail what exactly constitutes a violation of these usage policies, how potential violations are investigated, or how users who repeatedly abuse the service will be banned or otherwise sanctioned.

The NIST AI RMF generative AI companion and any updated NIST AI RMF must prioritize identifying the appropriate roles for developers and deployers and detailing risk mitigations for both developers and deployers. The initial NIST AI RMF did not significantly differentiate between developers and deployers, a significant omission that must be rectified moving forward.

Developers and deployers can make immediate and impactful changes to their current governance practices to manage the risks of generative AI. Below are some recommendations on risk mitigation from both the developers and the deployers drawn primarily from CAP's February 2024 report "Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone"[34] (the report is also attached in full at the end of these comments).

<u>The Role of Developers</u>

Below are ways that first-party developers can take immediate steps to shore up their systems in the short, medium, and long terms:

- **Incorporate specific recommendations for the implementation of the AI Bill of Rights for developers:** The NIST AI RMF generative AI companion and any updated NIST AI RMF should outline specific recommendations for how developers can implement the five principles from the AI Bill of Rights: (1) Safe and Effective Systems; (2) Algorithmic Discrimination Protections; (3) Data Privacy; (4) Notice and Explanation; and (5) Human Alternatives, Consideration, and Fallback.[35] These recommendations should include how developers can incorporate the principles of the AI Bill of Rights for model development and for first and third-party deployment of their AI models.

- **General Enforcement of Stated Policies**: Enforce all existing published policies. For example, if usage policies require disclosure of no "human in the loop" chatbot use cases, then an easily available method for reporting improper usage should be required of the deployer.
- **Reporting:** Anyone using an LLM at any time—in a first- or third-party capacity and in any format—should be able to report potential violations of an AI system to the developer and to the deployer of the LLM, if there is one, through a clear and transparent process with appropriate data retention. Reporting should be as frictionless as possible—for example, it should be displayed on the UI directly, and not require cold-emailing an address buried in a help-center article. Anthropic's updated reporting flow offers a strong example of this that other developers should imminently follow.[36]
- **Staffing**: Build and retain adequate internal staff for enforcement and maintenance of all policies, processes, and protocols to keep users safe.
- **API Access Oversight**: The developer should have a clear enforcement mechanism for managing API access and should build an enforcement regime to revoke access to third parties who violate usage policies, including with reporting, investigation, privacy-protecting documentation practices, appropriate data retention, and tooling to carry this out adequately. Build and enhance tooling to manage and revoke API access if a deployer violates developer terms or any other usage policies.
- **Require Additional Filtering**: Content moderation features—such as OpenAI's moderation endpoint[37] and Azure's abuse monitoring[38] —should be on by default for deployers using and manipulating developer LLMs; a submission of justification to the developer should be required before turning them off.
- **Data Access:** Retain access to inputs and outputs to ensure responsible system use and retain for an industry-agreed-upon amount of time before permanently deleting.
- **Abuse Prevention**: Developers should create tooling, such as content moderation endpoints and abuse monitoring, and make them easy for deployers to use and integrate into their apps.
- **Content Moderation Use Cases**: If an LLM is utilized for moderation of content generated by an LLM, that service should be provided for free or at a discount by the developer of the LLM, such as OpenAI's GPT-4 for content policy development and content moderation decisions.
- **Transparency**: Developers should be transparent with users about when they violate usage policies, including what actions or content led to the violations, how to appeal, and what remediations may be required of the user. Developers should also publish transparency reports for usage of LLMs to highlight prevalence of violations across abuse types and detailed reports from deployers of their technology.

<u>The Role of Deployers</u>

Below are ways that third-party deployers can take immediate steps to shore up their systems in the short, medium, and long terms:

- **Incorporate specific recommendations for the implementation of the AI Bill of Rights for deployers:** The NIST AI RMF generative AI companion and any updated NIST AI RMF should outline specific recommendations for how deployers of AI systems can implement the five principles from the AI Bill of

Rights: (1) Safe and Effective Systems; (2) Algorithmic Discrimination Protections; (3) Data Privacy; (4) Notice and Explanation; and (5) Human Alternatives, Consideration, and Fallback. These recommendations should include how deployers can incorporate the principles of the AI Bill of Rights in first and third-party deployment of AI models.

- **Data Sharing**: Ensure appropriate data-sharing mechanisms are in place between developers and deployers, with published retention policies for before, during, and after a report is made.
- **Reporting**: Deployers should, similarly to developers, be required to have a report function directly from the user to the deployer and to the developer, and should staff queues appropriately to ensure timely review against all relevant policies. Reporting should include a user-facing appeal flow and a commitment to human review appeals in a timely manner.
- **Staffing**: Build and retain adequate internal staff for enforcement and maintenance of all policies, processes, and protocols to keep users safe.
- **Transparency**: Deployers should be required to disclose which LLMs they are utilizing in their applications.

NIST should adopt the categories from the draft OMB AI guidance where AI use is presumed to be Safety-Impacting or Rights-Impacting and craft risk mitigations for these categories.

This RFI asks about the "Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics as defined in the AI RMF, as well as harms related to repression, interference with democratic processes and institutions, gender-based violence, and human rights abuses" and "Current standards or industry norms or practices for implementing AI RMF core functions for generative AI (govern, map, measure, manage), or gaps in those standards, norms, or practices."[39]

There is a critical lack of more granular standard categories that should be considered for AI risk management assessment and mitigation. Fortunately, there is a new body of work that has identified potential categories of AI use that should be leveraged in the NIST AI RMF generative AI companion and any updated NIST AI RM.

In November 2023, following the AI EO,[40] the Office of Management and Budget released a draft policy on "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence"[41] (draft OMB AI guidance). CAP provided extensive comments to OMB[42] on the draft OMB AI guidance[43] and joined the Leadership Conference,[44] Center for Democracy & Technology,[45] and others in their feedback.

Of particular note from the draft OMB AI guidance was Section 5.b, "Determining Which Artificial Intelligence Is Presumed to Be Safety-Impacting or Rights-Impacting."[46] This section outlines twenty-two categories in which the use of AI by government agencies should be "automatically presumed to be safety-impacting or rights-impacting."[47] These categories include critical infrastructure, law enforcement or surveillance, or emotional detection, among many others (categories automatically presumed to be safety-impacting or rights-impacting from the draft OMB AI guidance are attached for easy reference as Index 1).

While the draft OMB AI guidance was aimed at federal agencies using AI technologies, the list of categories where AI's use should be presumed safety-impacting or rights-impacting was a strong first official summation of risk categorization and represented a significant contribution to the broader Responsible AI canon.

NIST should adopt from (and build further upon) the draft OMB AI guidance the categories where AI usage are automatically presumed to be safety-impacting or rights-impacting and develop guidance on how to mitigate risk for each specific category in the Generative AI companion to the NIST AI RMF ordered by the AI EO and in future updates to the NIST AI RMF.

For example, the NIST AI RMF identifies three examples of potential harms from AI: Harm to People, Harm to an Organization, Harm to an Ecosystem.[48] Harms to People further identifies those potential harms to include "Individual: Harm to a person's civil liberties, rights, physical or psychological safety, or economic opportunity. Group/Community: Harm to a group such as discrimination against a population sub-group. Societal: Harm to democratic participation or education access."[49] But the current NIST AI RMF does not elaborate or detail any further categories for those harms to people, organizations, or ecosystems or how to mitigate for those specific harms.

The NIST AI RMF generative AI companion and any updated NIST AI RMF adopting the draft OMB AI guidance categories for AI usage that is presumed to be safety-impacting or rights- impacting and crafting specific risk mitigation guidance for those categories would prevent NIST from having to identify and craft its own list of categories for which to assess risk. It would also allow NIST to develop a common base of work with those developing compliance for federal agencies once the draft OMB AI guidance is finalized and allow private- and public-sector AI developers and deployers to benefit. An example of a category presumed to be safety-impacting or rights-impacting is detailed in the next section.

NIST should prioritize recommendations to address generative AI's risks to the integrity of elections and democratic processes given the historic number of elections taking place in 2024.

In 2024, more than 2 billion people will vote in more than 50 countries, including the United States, the European Union, and India.[50] AI-generated media is already impacting elections, as demonstrated by the 2023 Slovakian and Argentine elections in which AI was used to alter audio and generate images to damage candidates' reputations.[51] In the United States, generative AI has already made its debut in the presidential election process: in January, voters in New Hampshire received a robocall containing a deepfake of President Joe Biden's voice discouraging voters from voting in the primary.[52] Experts have raised significant concerns that generative AI is poised to impact elections in 2024 and beyond.[53]

Although not impossible, it will be a significant uphill battle for passage of federal legislation addressing the use of generative AI in elections in 2024. This means that a Generative AI companion to the NIST AI RMF addressing risk management with respect to elections and voting infrastructure could have an outsized impact on generative AI companies' approach to elections. Therefore, we urge NIST to make addressing generative AI's risks to the integrity of elections, including election security and administration, a top priority in the generative AI companion to the NIST AI RMF.

Given the broad range of categories where AI is presumed to be safety-impacting or rights-impacting in the draft OMB AI guidance, NIST is likely to have to prioritize the construction of detailed guidance for those areas. In urging NIST to prioritize risk mitigation guidance for the integrity of elections and democratic processes, we cite not only the 2 billion plus people heading to the polls next year, including the United States, but President Biden's prioritization of this issue. As noted above, CAP recommends NIST adopt the draft OMB AI guidance categories for AI usage presumed to be Safety-Impacting or Rights-Impacting.[54] In the draft OMB AI guidance, the first category where AI use should be presumed to be safety-impacting included "integrity of elections and voting infrastructure."[55] This is a clear indication of the importance of addressing the impact of AI, and generative AI in particular, on our elections and voting. This RFI specifically references "Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics as defined in the AI RMF, as well as harms related to repression, interference with democratic processes and institutions, gender-based violence, and human rights abuses."[56] The NIST AI RMF itself identifies a potential risk of AI for democratic participation, noting the potential harms related to AI systems include "Harms to People" and more specifically "Societal: Harm to democratic participation or education access." But it contains no further elaboration on the kinds of risks to democratic participation or ways to mitigate that risk.

First, in its Generative AI companion and subsequent revisions to the NIST AI RMF, NIST should outline a framework, in a style similar to its Privacy framework.[57] This framework should outline how generative AI developers and deployers, as well platforms and distributors, should mitigate both the creation and circulation of AI-generated content and prioritize the integrity of elections and democratic processes-related risks. This risk mitigation framework should focus on four key components:

1. Preventing AI systems from generating synthetic media that can impact the integrity of elections and democratic processes.
2. Responding to the generation of synthetic media that can impact the integrity of elections and democratic processes.
3. The distribution of synthetic media that can impact the integrity of elections and democratic processes.
4. Incorporating post-incident procedures to identify and adopt measures to prevent future synthetic media that can impact the integrity of elections and democratic processes.

The first prevention component must include specific recommendations on how to train models to not allow the generation of synthetic media that can impact the integrity of elections and democratic processes. This can be accomplished using Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF), safety tools at the input and output level, and other safety and responsible AI techniques for developers and deployers.[58] NIST should also consider recommendations for how to create standards for integrating provenance technologies into electoral systems—specifying in particular how digital content, such as voter information databases, election results, and voter education materials, can be tagged with provenance data to ensure accurate responses from generative AI tools.

Responding to the generation of synthetic media that can impact the integrity of elections and democratic processes requires addressing mitigation measures that generative AI developers and deployers can take to identify if their tools were used to generate synthetic content. This might require retaining logs or other supporting data, including the latest watermarking or attribution technology, as well as other responsible AI technologies to help identify synthetic media that can impact the integrity of elections and democratic processes. Such a framework should recommend that generative AI developers and deployers create general usage and elections-specific policies, staff teams to enforce those policies, be transparent about enforcement of usage policies, clearly outline first- and third-party enforcement around elections,[59] and block users and third-party deployers who misuse generative AI models.

Addressing the distribution of synthetic media that can impact the integrity of elections and democratic processes is a crucial component to holistic risk mitigation. The AI tools to generate synthetic content are only a part of the risk; we know that social media platforms are a major distributor of information online, and the advent of generative AI tools drastically reduces barriers to creating and disseminating mis- and disinformation throughout these platforms. This distribution of generative AI poses significant risks to an already weakened informational ecosystem and underscores the need for strong standards and governance of social media platforms. This should include general stipulations on how platforms must block the creation of harmful content to begin with, but also policies for dampening its distribution, reducing it in ranking systems, and deleting it altogether when detected.

Finally, the frameworks should develop recommendations for incorporating post-incident procedures to identify and adopt measures to prevent future synthetic media that can impact the integrity of elections and democratic processes. This could include the building tools, information sharing, and response to the distribution of those disruptive synthetic media. The lessons from any synthetic media that impact the integrity of elections or democratic processes incidents must be internalized and incorporated into the future prevention, response, and distribution steps described above. This should include recommendations on how developers can require deployers to report examples of synthetic media that can impact the integrity of elections and democratic processes, as well as for the developer to share those examples and other information broadly with all the deployers of their AI systems.

Second, NIST should differentiate between the components of the draft OMB AI guidance's safety-impacting category that includes the "integrity of elections and voting infrastructure"[60] to confer the highest level of risk management to safeguarding "voting infrastructure." Voting infrastructure and the legal and technical mechanics of election administration are the fundamental bedrock of our democracy and must be protected at all costs. Thus, they need the highest level of specificity and guidance to ensure both their security but also the public's trust. As CAP wrote last year: "While AI can improve the election process and strengthen the franchise, the potential for harm from AI in election administration is so great that it must have a highly compelling reason for its use. Any application of AI in elections should be treated as a component of election infrastructure and only be allowed after rigorous steps are taken to ensure it is safe, effective, transparent, auditable, and strengthens the franchise."[61] NIST is the component of the government best positioned to develop comprehensive testing standards for national laboratories to use when testing vendor or government-developed AI applications to be used in election infrastructure. NIST should call on (and partner with) the Election Assistance Commission (EAC) to develop voluntary guidelines that could be issued in the interim for any artificial intelligence systems proposed for use as part of election infrastructure.[62]

Another significant concern is the impersonation of election authorities through synthetic content, which can misguide voters about essential electoral procedures, such as voting methodologies, eligibility, polling station locations, and critical dates, thereby causing irreparable harm to the effective execution of democratic process. In addition, the ability of generative AI to write executable code poses a direct threat to the security of democratic processes by lowering the barrier to entry for creating sophisticated malware. This enables malicious actors, even those with minimal technical skills, to launch advanced cyberattacks against electoral systems and democratic institutions. There should be a clear focus on mitigations that can be rapidly deployed this year by generative AI companies given the number of elections in 2024, including those in the United States.

Third, the “integrity of elections”[63] and “interference with democratic processes and institutions”[64] encompasses a broader set of concerns around elections and democratic processes, including the potential for deep fakes of candidates and others to directly affect elections, the choices of voters, the general degradation of the information environment with AI-generated spam, and the specific use of electoral mis- and dis-information.

Fourth, the RFI asks for “Recommended changes for AI actors to make to their current governance practices to manage the risks of generative AI.” It is essential that the generative AI companion outline actual ways in which generative AI developers and deployers can and should identify risks to the integrity of elections and voting infrastructure and take concrete measures to prohibit and enforce those prohibitions. Many generative AI companies have usage policies that prohibit interfering in elections.[65] For example, the OpenAI usage policy states, “Don’t perform or facilitate the following activities that may significantly impair the safety, wellbeing, or rights of others, including: [. . .] Deterring people from participation in democratic processes, including misrepresenting voting processes or qualifications and discouraging voting.”[66] Yet OpenAI does not detail how it will enforce its usage policies (especially for third-party deployers accessing their models through APIs), does not require any additional safety features for third-party deployers accessing their models through APIs, and does not require any reporting from end users to developers about potential deployer abuses. For example, OpenAI also highlights examples of election violations in its January 2024 blog post “How OpenAI is approaching 2024 worldwide elections”[67] with four examples of violations, only one of which can be easily reported to OpenAI.

This RFI asks for “Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users)”[68] and, as noted above, it is essential that the generative AI companion and NIST AI RMF 2.0 identifies risks, responsibilities, and mitigations for both developers and deployers and how their usage policies are enforced for third-party usage. This is further detailed in CAP’s new report, “Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone.”[69]

Fifth, NIST should consider the AI recommendations set forth by CAP in our “Priorities for a National AI Strategy”[70] for inclusion in the NIST AI RMF generative AI companion and any updated NIST AI RMF including: enhancing cybersecurity measures for election infrastructure and ensuring AI system vendor security standards; updating aspects of the NSTC roadmap for information integrity research[71] to include generative AI mitigations; promoting and encouraging innovation to detect deep fakes and disinformation campaigns; and encouraging the limitation and responsibility of using AI systems as it pertains to election administration.

Conclusion

By making critical recommendations to manage the broad swath of risks posed by generative AI, NIST has a tremendous opportunity to shape the generative AI landscape in a positive and future-proof manner with the forthcoming AI RMF generative AI companion alongside any updates to the NIST AI RMF. NIST can play a crucial role in helping society balance harnessing the opportunities of generative AI while simultaneously mitigating its risk by incorporating the White House Blueprint for an AI Bill of Rights and adopting the categories from the draft OMB AI guidance where AI use is presumed to be Safety-Impacting or Rights-Impacting. Additionally, NIST should prioritize defining and outlining requirements for the responsibilities and risk management mechanisms for developers and first- and third-party deployers, and outline recommendations to address generative AI's risks to the integrity of elections and democratic processes given the historic number of elections taking place in 2024.

Index 1: Categories from the draft OMB AI guidance presumed to be Safety-Impacting or Rights-Impacting
For reference, below are the categories from Section 5.b from the draft OMB AI guidance, “Determining Which Artificial Intelligence Is Presumed to Be Safety-Impacting or Rights- Impacting,”[72] which we recommend NIST adopt for the NIST AI RMF generative AI companion and any updated NIST AI RMF:

i. **Purposes That Are Presumed to Be Safety-Impacting.** Unless the CAIO determines otherwise, covered AI within the scope of this memorandum is presumed to be safety-impacting and must follow the minimum practices for safety-impacting AI if it is used to control or meaningfully influence the outcomes of the following activities:

A. The functioning of dams, emergency services, electrical grids or the generation or movement of energy, fire safety systems, food safety mechanisms, integrity of elections and voting infrastructure, traffic control systems and other systems controlling physical transit, water and wastewater systems, and nuclear reactors, materials, and waste;
B. Physical movements, including in human-robot teaming, such as the movements of a robotic appendage or body, within a workplace, school, housing, transportation, medical, or law enforcement setting;
C. The application of kinetic force, delivery of biological or chemical agents, or delivery of potentially damaging electromagnetic impulses;
D. The movements of vehicles, whether on land, underground, at sea, in the air, or in space;
E. The transport, safety, design, or development of hazardous chemicals or biological entities or pathways;
F. Industrial emissions and environmental impact control processes;
G. The transportation or management of industrial waste or other controlled pollutants;
H. The design, construction, or testing of industrial equipment, systems, or structures
that, if they failed, would pose a meaningful risk to safety;
I. Responses to insider threats;
J. Access to or security of government facilities; or
K. Enforcement actions pursuant to sanctions, trade restrictions, or other controls on exports, investments, or shipping.

ii. **Purposes That Are Presumed to Be Rights-Impacting.** Unless the CAIO determines otherwise, covered AI is presumed to be rights-impacting (and potentially also safety- impacting) and agencies must follow the minimum practices for rights-impacting AI and safety-impacting AI if it is used to control or meaningfully influence the outcomes of any of the following activities or decisions:

A. Decisions to block, remove, hide, or limit the reach of protected speech;
B. Law enforcement or surveillance-related risk assessments about individuals, criminal recidivism prediction, offender prediction, predicting perpetrators' identities, victim prediction, crime forecasting, license plate readers, iris matching, facial matching, facial sketching, genetic facial reconstruction, social media monitoring, prison monitoring, forensic

analysis, forensic genetics, the conduct of cyber intrusions, physical location-monitoring devices, or decisions related to sentencing, parole, supervised release, probation, bail, pretrial release, or pretrial detention;

C. Deciding immigration, asylum, or detention status; providing risk assessments about individuals who intend to travel to, or have already entered, the U.S. or its territories; determining border access or access to Federal immigration related services through biometrics (e.g., facial matching) or other means (e.g., monitoring of social media or protected online speech); translating official communication to an individual in an immigration, asylum, detention, or border context; or immigration, asylum, or detention-related physical location- monitoring devices.
D. Detecting or measuring emotions, thought, or deception in humans;
E. In education, detecting student cheating or plagiarism, influencing admissions processes, monitoring students online or in virtual-reality, projecting student progress or outcomes, recommending disciplinary interventions, determining access to educational resources or programs, determining eligibility for student aid, or facilitating surveillance (whether online or in-person);
F. Tenant screening or controls, home valuation, mortgage underwriting, or determining access to or terms of home insurance;
G. Determining the terms and conditions of employment, including pre-employment screening, pay or promotion, performance management, hiring or termination, time-on-task tracking, virtual or augmented reality workplace training programs, or electronic workplace surveillance and management systems;
H. Decisions regarding medical devices, medical diagnostic tools, clinical diagnosis and determination of treatment, medical or insurance health-risk assessments, drug-addiction risk assessments and associated access systems, suicide or other violence risk assessment, mental-health status detection or prevention, systems that flag patients for interventions, public insurance care-allocation systems, or health-insurance cost and underwriting processes;
I. Loan-allocation processes, financial-system access determinations, credit scoring, determining who is subject to a financial audit, insurance processes including risk assessments, interest rate determinations, or financial systems that apply penalties (e.g., that can garnish wages or withhold tax returns);
J. Decisions regarding access to, eligibility for, or revocation of government benefits or services; allowing or denying access—through biometrics or other means (e.g., signature matching)—to IT systems for accessing services for benefits; detecting fraud; assigning penalties in the context of government benefits; or
K. Recommendations or decisions about child welfare, child custody, or whether a parent or guardian is suitable to gain or retain custody of a child.

Attachment 1: CAP Report: "Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone"

[1] Jon Porter, "ChatGPT continues to be one of the fastest-growing services ever," *The Verge*, November 6, 2023, available at https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer -conference.

[2] Executive Office of the President, "Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Press release, October 30, 2023, available at https://www.whitehouse.gov/briefingroom/presidential-actions/2023/10/30/executiveorder-on-the-safe-secure-and-trustworthydevelopment-and-use-of-artificial-intelligence/.

[3] Shalanda D. Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," Executive Office of the President, available at https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf (last accessed February 2024).

[4] Center for American Progress and others, "Letter-to-WH-on-AI-EO," August 3, 2023, available at https://www.americanprogress.org/wp-content/uploads/sites/2/2023/08/Letter-to-WH-on-AI-EO.pdf.

[5] Margot E. Kaminski, "Regulating the Risks of AI," *Boston University Law Review* 103 (5) (2023) 1347-1411, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4195066.

[6] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," (Washington: U.S. Department of Commerce, 2023) available at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[7] The White House, "Blueprint for an AI Bill of Rights," available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (last accessed February 2024).

[8] Megan Shahi and others, "Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone," (Washington, DC: Center for American Progress, 2024) available at https://www.americanprogress.org/article/generative-ai-should-be-developed-and-deployed-responsibly-at-every-level-for-everyone.

[9] Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence."

[10] Megan Shahi, "Protecting Democracy Online in 2024 and Beyond," Center for American Progress, September 14, 2023, available at https://www.americanprogress.org/article/protecting-democracy -online-in-2024-and-beyond/.

[11] The White House, "Blueprint for an AI Bill of Rights."

[12] Ibid.

[13] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)."

[14] Adam Conner, "The Needed Executive Actions to Address the Challenges of Artificial Intelligence," Center for American Progress, April 25, 2023, available at https://www.americanprogress.org/article/the-needed-executive-actions-to-address-thechallenges-of-artificial-intelligence/; and Megan Shahi and Adam Conner, "Priorities for a National AI Strategy," Center for American Progress, August 10, 2023, available at https://www.americanprogress.org/article /priorities-for-a-national-ai-strategy/.

[15] National Institute of Standards and Technology, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)," Federal Register 88 (244) (2023): 88368–88370, available at https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the.

[16] Shahi and others, "Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone."

[17] Ibid.

[18] National Institute of Standards and Technology, "developer," available at https://csrc.nist.gov/glossary/term/developer (last accessed January 2024).

[19] Sundar Pichai and Demis Hassabis, “Introducing Gemini: our largest and most capable AI model,” Press release, December 6, 2023, available at https://blog.google/technology/ai/google-gemini-ai/.
[20] National Institute of Standards and Technology, “Appendix A: Descriptions of AI Actor Tasks,” available at https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A (last accessed January 2024).
[21] Pichai and Hassabis, “Introducing Gemini: our largest and most capable AI model.”
[22] National Institute of Standards and Technology, “Appendix A: Descriptions of AI Actor Tasks.”
[23] Snapchat, “What is My AI on Snapchat, and how do I use it?”, available at https://help.snapchat.com /hc/en-gb/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it- (last accessed January 2024).
[24] Meta, “Introducing Llama 2.”
[25] MistralAI, “Frontier AI in your hands,” available at https://mistral.ai/ (last accessed February 2024).
[26] BigScience, “Introducing The World’s Largest Open Multilingual Language Model: BLOOM,” available at https://bigscience.huggingface.co/blog/bloom (last accessed January 2024).
[27] National Institute of Standards and Technology, “Glossary,” available at https://csrc.nist.gov/glossary (last accessed February 2024).
[28] Reuters, “OpenAI plans major updates to lure developers with lower costs, Reuters sources say,” CNBC, October 12, 2023, available at https://www.cnbc.com/2023/10/12/openai-plans-major-updates-to-lure- developers-with-lower-costs-reuters.html.
[29] Will Henshall, “4 Charts That Show Why AI Progress Is Unlikely to Slow Down,” Time, August 2, 2023, available at https://time.com/6300942/ai-progress-charts/.
[30] OpenAI, “Usage policies,” available at https://openai.com/policies/usage-policies (last accessed January 2024); Microsoft, “For Online Services,” available at https://www.microsoft.com/licensing/ terms/product/ForOnlineServices/all (last accessed February 2024); Microsoft, “Code of conduct for Azure OpenAI Service,” December 18, 2023, available at https://learn.microsoft.com/en-us/legal/cognitive-services/openai/code-of-conduct; Meta, “Llama Use Policy,” available at https://ai.meta.com/llama/use-policy/ (last accessed January 2024).
[31] OpenAI, “Usage policies.”
[32] Meta, “Llama Use Policy.”
[33] OpenAI, “Usage policies”; Microsoft, “For Online Services”; Microsoft, “Code of conduct for Azure OpenAI Service.”.”
[34] Shahi and others, “Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone.”
[35] The White House, “Blueprint for an AI Bill of Rights.”
[36] Anthropic, “How do I report harmful or illegal content?,” available at https://support.anthropic.com/en/articles/7996906-how-do-i-report-harmful-or-illegal-content (last accessed January 2024).
[37] OpenAI, “Moderation,” available at https://platform.openai.com/docs/guides/moderation (last accessed
January 2024).
[38] Microsoft, “Abuse Monitoring,” July 18, 2023, available at https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/abuse-monitoring.
[39] National Institute of Standards and Technology, “Request for Information (RFI) Related to NIST’s Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11).”
[40] Executive Office of the President, “Executive Order 14110: Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.”
[41] Young, “Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence.”

[42] Center for American Progress, "Re: Request for Comments: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Draft Memorandum; FR Doc. 2023-24269, 23 Nov. 2023.," December 5, 2023, available at https://www.americanprogress.org/wp-content/uploads/sites/ 2/2023/12/CAP-Draft-OMB-Comments-Final-12.04.2023.pdf.

[43] Executive Office of the President, "Request for Comments: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence Draft Memorandum," Federal Register 88 (212) (2023): 75625-75626, available at https://www.regulations.gov/document/OMB-2023-0020-0001.

[44] Leadership Conference On Civil and Human Rights and others, "Leadership Conference OMB AI Guidance Memo Comments," Leadership Conference on Civil and Human Rights, December 5, 2023, available at https://civilrights.org/resource/leadership-conference-omb-ai-guidance-memo-comments/.

[45] Ridhi Shetty and Alexandra Reeve Givens, "Civil Rights Organizations Identify Priorities for OMB Memo on Agency Use of AI," Center for Democracy and Technology, January 26, 2024, *available at* https://cdt.org/insights/civil-rights-organizations-identify-priorities-for-omb-memo-on-agency-use-of-ai/

[46] Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," p. 10.

[47] Ibid.

[48] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0), p. 5

[49] Ibid.

[50] Shahi, "Protecting Democracy Online in 2024 and Beyond."

[51] Morgan Meaker, "Slovakia's Election Deepfakes Show AI Is a Danger to Democracy," *Wired*, March 10, 2023, available at https://www.wired.co.uk/article/slovakia-election-deepfakes.
David Feliba, "How AI shaped Milei's path to Argentina presidency," The Japan Times, November 22, 2023, available at https://www.japantimes.co.jp/news/2023/11/22/world/politics/ai-javier-milei-argentina-presidency/.

[52] Ali Swenson and Will Wissert, "New Hampshire investigating fake Biden robocall meant to discourage voters ahead of primary," *Associated Press*, January 22, 2024, available at https://apnews.com/ article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5; and KAte Knibbs, "Researchers Say the Deepfake Biden Robocall Was Likely Made With Tools From AI Startup ElevenLabs," *Wired*, January 26, 2024, available at https://www.wired.com/story/biden-robocall -deepfake-elevenlabs/ .

[53] Anti-Defamation League and The Leadership Conference on Civil and Human Rights, "The Leadership Conference and ADL Comments to PCAST on Generative AI" (Washington, D.C.: The Leadership Council on Civil and Human Rights, 2023), available at https://civilrights.org/resource/leadership-conference-and-adl-comments-to-pcast-on-generative-ai/.
Norman Eisen and others, "AI can strengthen U.S. democracy—and weaken it" (Washington, D.C.: Brookings Institute, 2023), available at https://www.brookings.edu/articles/ai-can-strengthen-u-s-democracy-and-weaken-it/.
Rick Claypool and Cheyenne Hunt, "'Sorry in Advance!': Rapid Rush to Deploy Generative A.I. Risks a Wide Array of Automated Harms" (Washington, D.C.: Public Citizen, 2023), available at https://www.citizen.org/article/sorry-in-advance-generative-ai-artificial-intellligence-chatgpt-report/.

[54] Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence."

[55] Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," p. 11.
[56] National Institute of Standards and Technology, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)."
[57] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)."
[58] Meta, "Llama 2: Responsible Use Guide" (Menlo Park, CA: 2023), available at https://ai.meta.com/static-resource/responsible-use-guide/; OpenAI, "GPT-4 System Card" (San Francisco: 2023), available at https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[59] Shahi, "Protecting Democracy Online in 2024 and Beyond."
[60] Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," p. 11.
[61] Shahi and Conner, "Priorities for a National AI Strategy."
[62] Shahi and Conner, "Priorities for a National AI Strategy."
[63] Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," p. 11.
[64] National Institute of Standards and Technology, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)."
[65] Microsoft, "Code of conduct for Azure OpenAI Service," and OpenAI, "Usage policies,"
[66] OpenAI, "Usage policies."
[67] OpenAI, "How OpenAI is approaching 2024 worldwide elections," January 15, 2024, available at https://openai.com/blog/how-openai-isapproaching-2024-worldwide-elections.
[68] National Institute of Standards and Technology, "Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)."
[69] Shahi and others, "Generative AI Should Be Developed and Deployed Responsibly at Every Level for Everyone."
[70] Shahi and Conner, "Priorities for a National AI Strategy."
[71] National Science and Technology Council, "Roadmap for Researchers on Priorities Related to Information Integrity Research and Development," (Washington, DC: 2022) available at https://www.whitehouse.gov/wp-content/uploads/2022/12/Roadmap-Information-Integrity-RD-2022.pdf.
[72] Young, "Proposed Memorandum For The Heads Of Executive Departments And Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence."