# Response to NIST Request for Information (RFI) Related to the AI Executive Order

1 February 2024

Elham Tabassi, Chief of Staff, Information Technology Laboratory
ATTN: AI E.O. RFI Comments
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 8900, Gaithersburg, MD 20899–8900

Subject: NIST AI Executive Order

Via email to ai-inquiries@nist.gov

To Ms. Tabassi, and the entire NIST team carrying out responsibilities under the AI Executive Order,

Thank you for the invitation to submit comments in 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). We offer the following submission for your consideration.

We are researchers affiliated with UC Berkeley, with expertise on AI research and development, safety, security, policy, and ethics. We previously submitted responses to NIST several times over the past two and a half years at various stages of NIST's development of the AI Risk Management Framework (AI RMF) and follow-on work such as NIST's Generative AI Public Working Group.

In the following comments, we aim to say the most about RFI issues that may have received relatively little attention to date in NIST forums such as the Generative AI Public Working Group, e.g., "Different risk profiles and considerations for synthetic content for models with widely available model weights" as listed under RFI topic 2a. We also aim to follow the RFI's guidance to provide "information that is specific and actionable" rather than "general statements about the challenges and needs".

Thank you again for the opportunity to comment on the NIST RFI related to the AI Executive Order. If you need additional information or would like to discuss further, please contact Anthony Barrett at anthony.barrett@berkeley.edu or Jessica Newman at jessica.newman@berkeley.edu.

In any case, we look forward to further engagement with NIST as you and others act on responsibilities under the AI Executive Order.

Our best,

Anthony Barrett, Ph.D., PMP
Visiting Scholar
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Jessica Newman
Director, AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
Co-Director, AI Policy Hub, UC Berkeley

Brandie Nonnecke, Ph.D.
Director, CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley
Co-Director, AI Policy Hub, UC Berkeley
Assoc. Research Professor, Goldman School of Public Policy, UC Berkeley

Evan R. Murphy
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Krystal Jackson
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

# Our comments on specific items in the NIST RFI related to the AI Executive order

In the following, we list NIST RFI questions for which we provide answers, and omit NIST RFI questions that we do not specifically address.

## 1. Developing Guidelines, Standards, and Best Practices for AI Safety and Security

### a. (1) Developing a companion resource to the AI RMF for generative AI

One resource that provides guidance and links to resources for identifying impacts of generative AI systems and mitigations for negative impacts is our own November 2023 publication, the UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models, Version 1.0 (see https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile/). For over a year, we led development and testing of the profile, with input and feedback from more than 100 people representing a range of stakeholders, resulting in over 100

pages of guidance and accompanying material for developers of cutting edge GPAIS and foundation models. The profile is aligned with the NIST AI Risk Management Framework (AI RMF) and other AI standards such as ISO/IEC 23894. The Berkeley GPAIS and foundation model profile effort is separate from, but aims to complement and inform the work of, other guidance development efforts such as the NIST Generative AI Public Working Group.

The Berkeley GPAIS and foundation model profile discusses a wide variety of risks and harms of generative AI, highlights different roles for different AI actors (e.g., the role of AI developers vs. deployers), and discusses current techniques and implementations for managing risks and harms of generative AI, including the importance of documentation, reporting, and engagement.

## 2. Reducing the Risk of Synthetic Content, Topics under subsection (a):

### Preventing generative AI from producing child sexual abuse material or producing non-consensual intimate imagery of real individuals (to include intimate digital depictions of the body or body parts of an identifiable individual); and Ability for malign actors to circumvent such techniques

According to the 2023 State of Deepfakes (see https://www.homesecurityheroes.com/state-of-deepfakes/) deepfake pornography makes up 98% of all deepfake videos online and 99% of the individuals targeted in deepfake pornography are women. These videos can be made for free in less than 25 minutes. The harms of non-consensual intimate imagery disproportionately impact women and girls of particular racial, ethnic, and religious backgrounds (see https://unesdoc.unesco.org/ark:/48223/pf0000387483/PDF/387483eng.pdf.multi). The UK Online Safety Act of 2023 prohibits the sharing of non-consensual deepfake pornography, but the US does not yet have a federal law to protect women and children from the harms of AI generated non-consensual intimate imagery. Unfortunately leading technical solutions such as watermarking, labeling, and authenticating provenance will do little to stop this, and so it will be particularly critical to support robust content moderation across media platforms to facilitate the rapid removal of exploitative and illegal content and to provide redress for those harmed, in addition to criminalizing the creation and intentional spreading of non-consensual intimate imagery including child sexual abuse material (CSAM).

### Different risk profiles and considerations for synthetic content for models with widely available model weights

Foundation model developers that publicly release the model parameter weights for their models with downloadable, fully open, or open source access to their models, and other foundation model developers that suffer a leak of model weights, will in effect be unable to shut down or decommission AI systems that others build using those model weights. Moreover, direct access to model weights can also make it easier for malicious actors to remove or otherwise circumvent safeguards that a foundation model built into the original foundation model for the model's release. These are considerations that should be weighed against the benefits of models with widely available parameter weights, especially for the largest-scale and most broadly capable models that pose the greatest risks of enabling severe harms, including from

malicious misuse to harm the public. Many of the benefits of open source models, such as review and evaluation from a broader set of stakeholders, can be supported through transparency, engagement, and other openness mechanisms that do not require making a model's parameter weights downloadable or open source, or by releasing smaller-scale and less broadly capable open source models.

Foundation model developers that plan to provide downloadable, fully open, or open source access to their models should first use a staged-release approach (e.g., not releasing parameter weights until after an initial closed source or structured access release where no substantial risks or harms have emerged over a sufficient time period), and should not proceed to a final step of releasing model parameter weights until a sufficient level of confidence in risk management has been established, including for safety risks and risks of misuse and abuse. (The largest-scale or most capable models should be given the greatest duration and depth of pre-release evaluations, as they are the most likely to have dangerous capabilities or vulnerabilities that can take some time to discover.)

We provide the above guidance, and related material, under Manage 2.4 of our AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models, Version 1.0 (see https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile/).

## 3. Advance Responsible Global Technical Standards for AI Development, Topics under subsection (a):

### AI systems for which standards would be particularly impactful (e.g., because they are especially likely to be deployed or distributed across jurisdictional lines, or to need special governance practices)

Many regulatory requirements focus on specific industry sectors and end-use applications, e.g., in critical infrastructure or other high-risk categories of the draft EU AI Act. While valuable for downstream developers of end-use applications, an approach focused on end-use applications could overlook an opportunity to provide profile guidance for upstream developers of foundation models. Such AI systems can have many uses, and early-development risk issues such as emergent properties that upstream developers are often in a better position to address than downstream developers building on AI platforms for specific end-use applications.

### Guidelines and standards for trustworthiness, verification, and assurance of AI systems

In our research, documented for example in A Taxonomy of Trustworthiness for Artificial Intelligence (see: https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence-standalone-taxonomy/) we have found that many considerations related to the safe and trustworthy development of AI systems are best addressed during the early design and development stages of the AI lifecycle. Solely focusing on end-use and on the deployers of AI systems misses the importance for example of standards for privacy and security by design,

for data quality and curation, and for testing and evaluation of general capabilities and vulnerabilities.