PARTNERSHIP ON AI

The Partnership on AI response to the NTIA Request for Comment (RFC) on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

Background

Partnership on AI (PAI) is a non-profit partnership of academic, civil society, industry, and media organizations creating solutions to ensure that AI advances positive outcomes for people and society. PAI studies and formulates sociotechnical approaches aimed at achieving the responsible development of artificial intelligence (AI) and machine learning (ML) technologies. Today, we connect over 100 partner organizations in 14 countries to be a uniting force for the responsible development and fielding of AI technologies.

PAI develops tools, recommendations, and other resources by inviting multistakeholder voices from across the AI community and beyond to share insights that can be synthesized into actionable guidance. We then work to promote adoption in practice, inform public policy, and advance public understanding. We are not an industry or trade group nor an advocacy organization. We aim to change practice, inform policy, and advance understanding.

The information in this document is provided by PAI and is not intended to reflect the view of any particular Partner organization of PAI. The comments provided herein are intended to provide evidence-based information, based on PAI's research, in response to NTIA's RFC.

Executive Summary

PAI welcomes the opportunity to provide this input to NTIA's work under section 4.6 of the Executive Order on "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (the "AI Executive Order").

This submission responds to NTIA's request for comment (RFC) on:

- Risks and benefits of dual use foundation models with widely available model weights
- Potential voluntary, regulatory, and international mechanisms to manage the risks and maximize the benefits of these models.

This paper adopts the RFC's usage of "open foundation models" to refer to models with widely available model weights, while noting that model weights are not the only component of models that can be released by providers.[1]

As the RFC notes, open foundation models promise a number of benefits. However, all models — whether they are open or more closed — present risks. It is vital that appropriate steps are taken to address these risks while promoting benefits. To get this balance right, any mechanisms for risk mitigation for foundation models should be evidence-based, and be appropriately tailored to target identified risks, including by taking into account model capabilities and release type. Applying these principles to open foundation models, specific risk mitigation measures should only be imposed on open foundation models when that is necessary to address some differential or **marginal risk** posed by those models.

This submission draws on work undertaken by PAI in preparing its Guidance for Safe Foundation Model Deployment (the "Model Deployment Guidance"). The Guidance is a voluntary framework for foundation model providers, containing detailed recommendations to identify and mitigate risks associated with these models.[2] It provides specialized guidance tailored to different model capabilities, and to different model release types — including releases where model weights are made widely available at the time of deployment.

PAI submits that NTIA's work under the AI Executive Order should be informed by the following principles:

- **All foundation models need risk mitigations.** Appropriate risk identification and mitigation practices should be adopted for the

[1] In PAI's recent Guidance for Safe Foundation Model Deployment, we use the term "open access" release, to refer to models released publicly with full access to at least model weights.
[2] PAI's Model Deployment Guidance defines **foundation models** "to encompass all models with generally applicable functions that are designed to be used across a variety of contexts. The current generation of these systems is characterized by training deep learning models on large datasets (which requires significant computational resources) to perform numerous tasks that can serve as the "foundation" for a wide array of downstream applications." **Model providers** are defined to be those training foundational models (proprietary or open-source) that others may build on as well as interfaces to interact with the models.

development and deployment of **all** foundation models — from narrow specialized models to frontier models, and from fully closed models to fully open models.

- **Appropriate risk mitigations will vary depending on model characteristics.** PAI's Model Deployment Guidance tailors its recommendations according to model capability and release type.
- **Risk mitigation measures, for either open or closed models, should be proportionate to risk**. Recommended risk management practices for both open and closed foundation models should be proportionate to the risks posed, should be based upon the best evidence available, and should be updated to reflect new evidence as it emerges.
- **Voluntary frameworks are part of the solution.** Existing voluntary risk management frameworks, such as PAI's Model Deployment Guidance, play a valuable role in providing benchmarks for safe foundation model deployment, including when it is proposed to make model weights widely available, and can continue to do so as the policy landscape evolves.

Research on the benefits and risks of open foundation models, and appropriate risk mitigation measures, is ongoing. This year, PAI will be continuing its work on this topic consulting with its partners to explore these issues and identify current and emerging best practices. **PAI would be happy to update NTIA as this work progresses, and to contribute to future stages of NTIA's work on this issue.**

Summary of Recommendations

1. In fulfilling its tasks under section 4.6 of the AI Executive Order, NTIA should consult widely with all relevant stakeholders from civil society, academia, and private industry, ensuring all have equal opportunity to contribute. Diverse perspectives are essential to inform the NTIA on all relevant issues including the benefits of open models and risks. Multistakeholder forums such as PAI play an important role in facilitating cross-sectoral participation. PAI would be happy to assist NTIA in this process.

2. NTIA should promote the role and adoption of voluntary risk management frameworks such as PAI's Model Deployment Guidance by providers of both open and closed foundation models. NTIA should consider the role these frameworks play in addressing risk right now, and when assessing other potential policy or regulatory responses in fulfilling its mandate under the AI Executive Order.

3. NTIA should recommend the introduction of mechanisms restricting the deployment and use of foundation models (whether open or closed) only where that is necessary and proportionate to mitigating risk, on the basis of the best available evidence.

4. NTIA should recommend the introduction of mechanisms specifically restricting the deployment and use of open foundation models only where, on the best available evidence, that is necessary and proportionate to mitigating risks specifically associated with the wide availability of the model's weights. That is, open foundation models should not be subject to special restrictions unless they pose a higher risk than closed models, or pose some risk that closed models do not.

Open foundation models/"models with widely available model weights"

RFC questions 1-3[3]

As recognized in the AI Executive Order, AI brings the promise of great benefits, promoting innovation and productivity, and helping solve today's most pressing challenges. It can also present significant risks.[4] These observations are particularly relevant for foundation models. Responsible foundation model deployment requires that providers implement appropriate mechanisms to identify and mitigate risk across the model lifecycle. PAI's Model Deployment Guidance represents our most recent work in this area, containing recommendations to assist providers operationalize AI safety principles. The Guidance reflects the fact that **risk mitigations are necessary for all foundation models for all release types, and all model capabilities**. Mitigations should also be scaled to according to model characteristics, to ensure that higher risk models are subject to additional safeguards.

Risks and Benefits of Open Access Models

As acknowledged in the RFC, open foundation models can lead to a range of benefits, including by increasing access to models, enabling research, and promoting transparency. PAI supports policy settings that enable responsible open access model deployments, as reflected by its endorsement of the recent open letter led by PAI partners, the Center for Democracy and Technology and Mozilla. Consistently with that letter, restrictions or specific safety guidance for open foundation models are only warranted when those models present specific risks that are greater or different from those presented by closed models. This is consistent with the approach taken in PAI's Model Deployment Guidance — which generates custom risk guidance for deployers that is scaled to model risk based on both the capabilities and the release type of a particular model.

---

[3]Question 1 of the RFC: *How should NTIA define "open" or "widely available" when thinking about foundation models and model weights?*
Question 2 of the RFC: *How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?*
Question 3 of the RFC: *What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?*

[4] AI Executive Order, section 1.

Marginal risk of open foundation models

Like closed foundation models, open access models can present risks.[5] A recent case study from PAI's AI and Media Integrity program records the use of open foundation models in a global elections context, and notes the limitations of downstream mitigation measures for synthetic media harm that rely solely on responsible practices by "good actors." However, this does not mean such mitigations lack value. The level of friction these mitigation measures are able to create to slow the spread of harmful synthetic media makes them still worth implementing.

In assessing the risk posed by open foundation models, and appropriate measures to address those risks, policy makers should focus on the marginal risks associated with open access release.[6] This approach reflects the fact that both open and closed models present risks (and warrant mitigations to address those risks). Risk mitigation measures that **specifically target open models**, or impose differential restrictions on open models, should only be recommended **where there is evidence that open models may pose risks beyond those posed by closed models**.

Some of the features of open models that may be relevant to assessing differential risk include that open release of model weights is irreversible, and that moderation/monitoring of open models post-release is challenging. These factors are particularly relevant in the context of frontier models. PAI's Deployment Guidance defines "frontier" or "paradigm-shifting" models as:

> "Cutting edge general purpose models that significantly advance capabilities across modalities compared to the current state of the art."[7]

In assessing whether a model meets this definition, considerations include:

- Does the model enable significantly more advanced capabilities compared to current state-of-the-art?

[5] Foundation models (both open and more closed) can give rise to a variety of risks, including risks arising from models themselves and also from downstream use. PAI's Model Deployment Guidance includes a mapping of in-scope foundation model risks: https://partnershiponai.org/modeldeployment/#learn_more.
[6] PAI's Model Deployment Guidance uses the terminology "open access release" to refer to models where at least model weights are released at deployment. Other model components - such as data and code - might also be released. There is substantial overlap between this concept and that of "models with widely available model weights."
[7] Partnership on AI, Guidance for Safe Foundation Model Deployment (2023).

- Does the model utilize parameters or computational resources that greatly exceed current standards, demonstrating a breakthrough in scalable training?
- Does the model show evidence of self-learning capabilities exceeding current AI?
- Does the model provider enable execution of commands, or actions directly in the real world through released interfaces or applications, beyond passive information processing?

Frontier models present some particular safety challenges. Capabilities of these models, including dangerous capabilities, can emerge unexpectedly. They can therefore present more "unknown unknowns".

> "[I]t is difficult to robustly prevent a deployed model from being misused; and, it is difficult to stop a model's capabilities from proliferating broadly."

For this reason, PAI's Model Deployment Guidance recommends that frontier or paradigm-shifting models not be subject to open access release without first probing for those risks, and assessing the effectiveness of risk mitigations. As the Guidance states:

> "We recommend providers initially err towards staged rollouts and restricted access to establish confidence in risk management for these systems before considering open availability.
>
> "These models may possess unprecedented capabilities and modalities not yet sufficiently tested in use, carrying uncertainties around risks of misuse and societal impacts. Over time, as practices and norms mature, open access may become viable if adequate safeguards are demonstrated."[7]

This is consistent with the principles set out at the beginning of this section — that appropriate risk mitigations should be adopted for all foundation models prior to deployment, and that those mitigations should be appropriately tailored to respond to relevant risks.

Recommendations to NTIA for open foundation models

In considering what policy mechanisms are appropriate for open foundation models, PAI submits that the NTIA:

- Should recommend imposing restrictions on foundation models (whether open or closed) only where that is demonstrated to be necessary and proportionate to mitigate risk
- Should recommend that specific restrictions be placed on open access models only when that is demonstrated to be necessary and proportionate to mitigating a category of risk that is specifically associated with the open access release of the model.

PAI will be continuing its work on open access models this year, including through a workshop in early April to explore the need for and effectiveness of risk management strategies for the open foundation model ecosystem; to assess current guidance for open access release of state-of-the-art models; and to consider the effectiveness of current and potential new harm reduction strategies. **PAI will update NTIA as this work progresses, and contribute to future stages of NTIA's work on this issue.**

PAI's Model Deployment Guidance

(RFC question 7)[8]

The RFC calls for submissions addressing current and potential mechanisms to manage the risks and maximize the benefits of open foundation models. One such mechanism is PAI's Guidance for Safe Foundation Model Deployment (the "Model Deployment Guidance").

PAI's Model Deployment Guidance is a voluntary set of practices intended to help foundation model providers identify and mitigate model risks, and inform public policy and understanding. PAI developed the Guidance working with stakeholders from over 40 global institutions, including model providers, civil society organizations, and academic institutions.

The Guidance contains 22 recommendations for safe foundation model deployment practices across the product lifecycle, from research and development through to post-deployment monitoring and decommissioning.

A key feature of the Guidance is that it allows developers to generate customized guidelines that are tailored to two key model attributes: model capability and release type. The recommended guidelines are more extensive for more capable models and more available release types. The Guidance defines:

[8] Question 7 of the RFC: *What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the risks and maximize the benefits of foundation models with widely available weights? What kind of entities should take a leadership role across which features of governance?*

- **Four categories of model release:** Open Access; Restricted API and Hosted Access; Closed Development; and Research Release.
- **Three categories of model capability:** Specialized Narrow Purpose; Advanced Narrow and General Purpose; and Paradigm-shifting or Frontier.

This approach allows the Guidance to provide risk management guidelines that are appropriately adapted to the risk profile of particular models.

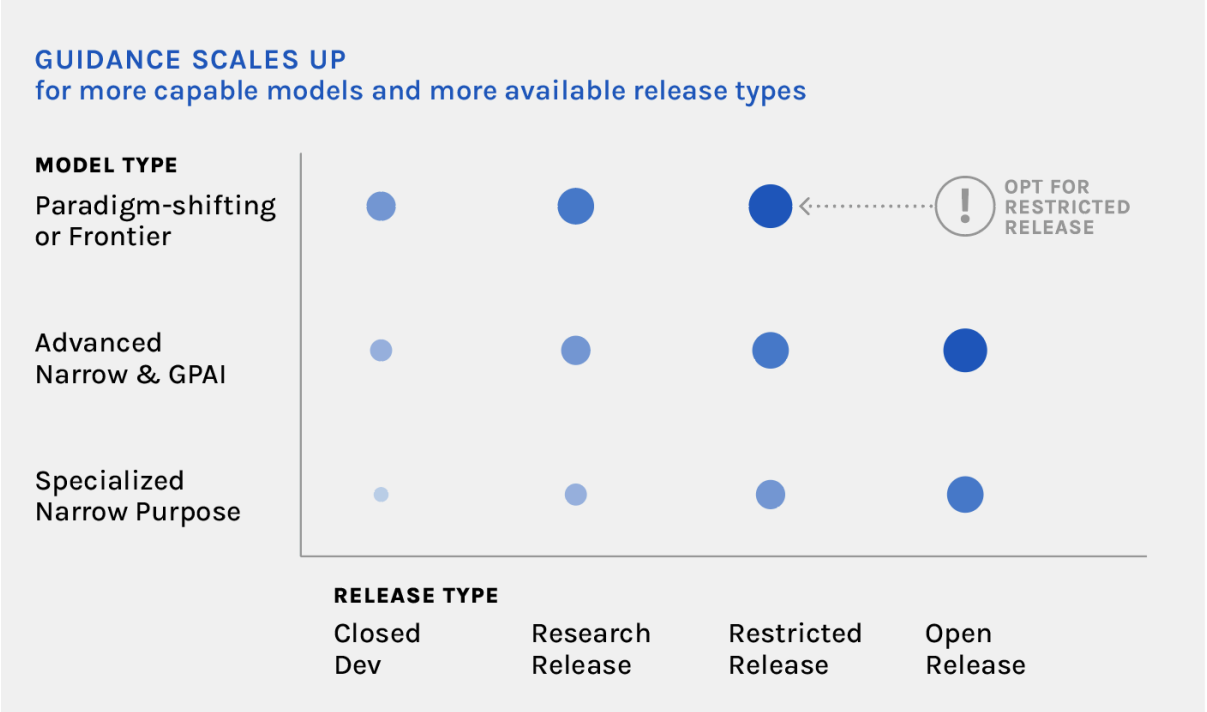

Figure 1: visualization of the scaled approach to safety guidelines under the Model Deployment Guidance

A full list of the 22 guidelines can be found here; further information about the Model Deployment Guidance can be found here; and customized Guidance can be generated here.

Relevant to NTIA's current work:

- The Model Deployment Guidance contains safety recommendations for the development/deployment of **both open and closed models**, that is adapted to the risks and appropriate mitigations for each of those categories. This approach reduces the compliance burden on providers of less risky models, while ensuring that riskier models are subjected to more stringent safety practices. The approach in the Guidance reflects the fact that model release type is one of a number of relevant factors in assessing what risk management measures should be implemented by model providers to

ensure safe deployment. (The Guidance uses the term "open access" to refer to model releases involving the release of at least model weights.)[9]

- **The Model Deployment Guidance can generate guidance for foundation models of all capabilities**, including "dual-use foundation models".[10]

The Model Deployment Guidance states the following:

- Risk management practices are necessary **for all foundation models.**
- The risks posed by foundation models vary, **depending on model capability and release type** (including open access release).
- Risk management measures for foundation models should be flexible, so **they are appropriate and adapted to the risks posed by those models**.

Recommendations developed by NTIA should reflect these principles.

As explained in the supporting materials for the Model Deployment Guidance, the Guidance was not developed to replace regulation or other policy frameworks. Rather, the Guidance aims to complement other governance and regulatory approaches. In the absence of policy frameworks, voluntary mechanisms like the Guidance play a valuable role in providing recommendations for risk management practices that can be adopted right now by a wide range of model providers. This helps ensure that benefits from these models, including open foundation models, are harnessed while risks are managed. Voluntary frameworks play a particularly important role in setting a benchmark for safe practice while work continues to better understand the risks associated with various types of model release.

We urge NTIA to encourage the adoption of voluntary risk management frameworks such as PAI's Model Deployment Guidance by providers of both open and closed foundation models. NTIA should consider the role these frameworks play in addressing risk right now and providing an initial benchmark for other potential policy or regulatory responses in fulfilling its mandate under the AI Executive Order.

---

[9] While noting that the definition of "widely available model weights" is one factor NTIA is considering under the AI Executive Order, there is likely to be significant overlap between the category of models with "widely available model weights" and that of models subject to "open access release".

[10] The RFC mirrors the AI Executive Order in considering risk mitigation for dual-use foundation models; it gives a definition of those models, including an interim definition for model reporting that is based on compute. There is some discussion in the literature (see also here) about how well this definition tracks frontier model capabilities. PAI's Model Deployment Guidance does not specifically address "dual-use foundation models" as a separate category; however because the Guidance generates customized recommendations for models of all capabilities, it can generate guidance for dual-use models.

Conclusion

PAI would be happy to provide further information about any of the matters discussed in this submission. We look forward to the NTIA's pending report.

For any further information and questions related to this submission, please contact John@partnershiponai.org (and copy policy@partnershiponai.org).