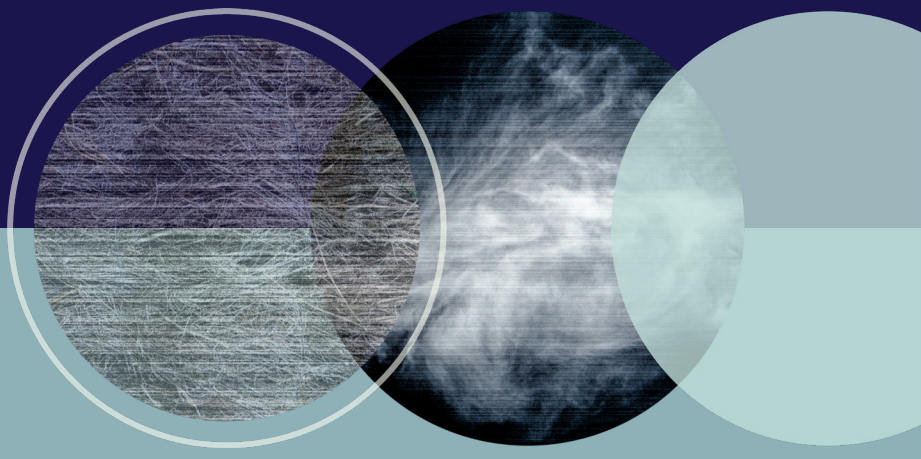
Policy and Governance
Working Group

March
2024

Recommendations to the US Department of Commerce (NTIA) on Open Foundation AI Models



SCIENCE, TECHNOLOGY,
AND SOCIAL VALUES LAB
INSTITUTE FOR ADVANCED STUDY

March 27, 2024

National Telecommunications and Information Administration
1401 Constitution Avenue NW
Washington, DC 20230
submitted via https://www.regulations.gov

Comment of the AI Policy and Governance Working Group
on the NTIA Request for Comment
on Dual Use Foundation Artificial Intelligence Models
with Widely Available Model Weights
Docket NTIA-240216-0052

The Biden-Harris Administration is soliciting insights about the "potential risks, benefits, other implications, and appropriate policy and regulatory approaches to dual-use foundation models for which the model weights are widely available" via the Department of Commerce's National Telecommunications and Information Administration (NTIA) request for comment. We commend the Administration's commitment to public consultation on an issue so important to the safe, secure, and trustworthy advancement, development, and use of artificial intelligence (AI).

Openly available data, code, and infrastructure have been critical to the advancement of science, technological innovation, economic growth, and democratic governance. These open resources have been built and shared in the context of commitments to open science, to expanding industry and markets, and to the principle that some technologies should be widely available for maximum public benefit, while allowing for control of access to data, code, and infrastructure as necessary for safety and security purposes. We recommend that the Biden-Harris Administration take a similarly measured approach to the governance of AI, including dual-use foundation models with widely available model weights, referred to in this comment as "open foundation models."

Open foundation models offer an avenue to achieve many of the United States' policy goals. As AI governance is developing globally, policy solutions related to people's rights and safety have tended to focus on increasing transparency and accessibility of AI systems to improve accountability of AI developers and deployers to the public. But beyond that, a robust open foundation model ecosystem is crucial to enabling a diverse, innovative, and competitive environment for technological innovation, as well as to

addressing concerns about any single entity accumulating excessive sector influence.[1] If well executed, an open foundation model ecosystem will expand collective understanding of AI beyond those who currently build and release models, spurring ingenuity and enabling a powerful base for innovation for researchers and developers across a range of sectors.

While *closed* foundation models potentially offer the possibility of risk monitoring and mitigation from developers and deployers, there are still many challenges to overcome to realize these benefits in practice. By contrast, It is crucial to consider that the broad and wide distribution of open foundation models may amplify myriad risks.[2] [3] [4] [5] Indeed, by their very nature, open foundation models released immediately or rapidly after training reduce developers' ability to monitor for and safeguard against misuse, and also make it more difficult to identify and hold accountable those responsible for misuse. As foundation model capabilities develop, the release of powerful AI models could create risks that are difficult to foresee, compounding existing challenges around accountability for developers and deployers, and making them more difficult to address.

When reviewing the risks, benefits, and implications of dual-use open foundation models, we believe the U.S. government must take two issues into account:

- First, mechanisms and strategies for model release and model access exist along the spectrum between the extreme poles of fully open and fully closed models.[6] Any analysis of “marginal risks”[7] of open foundation models should similarly take place across this spectrum of known and unforeseeable use cases.
- Second, components of the foundation model stack can be accessed and modified through a spectrum of staged and structured approaches that blur the

[1] Vipra, J., & Korinek, A. (2023). Market concentration implications of foundation models. *arXiv preprint arXiv:2311.01550*.
[2] Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., ... & Dafoe, A. (2023). Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.
[3] Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., ... & Gupta, A. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*.
[4] Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., ... & Narayanan, A. (2024). On the Societal Impact of Open Foundation Models.
[5] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
[6] Solaiman, I. (2023, June). The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 111-122).
[7] Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., ... & Narayanan, A. (2024). On the Societal Impact of Open Foundation Models. Marginal risk refers to “the extent to which these models increase societal risk by intentional misuse beyond closed foundation models or pre-existing technologies, such as web search on the internet.”

> binary between open and closed.[8] When fully open access to AI models is not possible, these approaches may offer a strategy to provide controlled but greater access to otherwise-closed components of AI models (e.g., the ability to perform fine-tuning on a proprietary model in a controlled setting).

As such, the idea that model weights are the fulcrum for reasoning about the societal impact of foundation models is not universally applicable. In some cases, factors such as model development and nature of deployment may be far more relevant or revelatory. In other instances in which access to open resources is particularly relevant, the release status of assets beyond weights (e.g., training and fine-tuning data, training and inference code, etc.) should also shape government's analysis. Indeed, the focus on the term "model weights" may be misleading in that it obscures other processes and data flows that are crucial in the actual deployments and uses of open foundation models and, therefore, for the fuller understanding of their risks and benefits.

A fuller consideration will require:

> 1) clarity about the primary national goals of broad access to dual-use, open foundation model AI,
>
> 2) mandated developer disclosure combined with agility in determining "thresholds" and other signals of interest, as AI model capabilities evolve, and
>
> 3) an exploration of the "spectrum of access," that is, frameworks that move past binary release to alternatives such as staged release, [9] structured access,[10] and other frameworks that support these national goals.

The AI Policy and Governance Working Group recommends that the Biden-Harris Administration call for the development of a range of practical approaches to open foundation model release with accompanying case studies and pilot studies developed with relevant stakeholders.

Alongside this recommendation, we urge *precautionary friction* in which policymakers embrace small delays and testing appropriately calibrated to the risk of release, rather than wholesale restrictions on open foundation models that have a broad range of potential beneficial uses. But we also argue that this friction should be accompanied by strong policy bias towards supporting the appropriate accessibility and availability of

---

[8] Bluemke, E., Collins, T., Garfinkel, B., & Trask, A. (2023). Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases. *arXiv preprint arXiv:2303.08956*.
[9] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
[10] Bucknall, B. S., & Trager, R. F. (2023). Structured Access For Third-Party Research On Frontier AI Models: Investigating Researchers' Model Access Requirements.

foundation models, allowing transparency into models for both accountability and discovery, and encouraging innovation beyond a few organizations with extensive computing infrastructure.

AI Policy and Governance Working Group

Members of the AI Policy and Governance Working Group represent a mix of sectors, disciplines, perspectives, and approaches. Despite these differences, we agree that it is not only possible but necessary to address the multitude of concerns raised by the expanding use of AI systems and tools and their increasing capabilities. We also agree that both present-day harms and risks on the horizon posed by various AI models warrant urgent attention in order to fulfill the public's legitimate expectation of safety and respect for their rights. We share the belief that these issues require immediate and ongoing action from industry, governments, academia, and civil society to meet public expectations. To this end, we have previously submitted recommendations to the NTIA in response to its request for comment on algorithmic accountability and to the United Nations Secretary-General's Office of the Technology Envoy in response to its request for expert advice on the global governance of AI.[11]

Open Foundation Models Contribute to US Policy Goals

As demonstrated by President Biden's October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, the Biden-Harris Administration appreciates that AI may have broad impact and that governance of AI is important to many of its key policy objectives. At its best, open access to technology can help to advance innovation, reduce concentrations of expertise and power, enable transparency, and create new avenues to ensure societal safety.

Access to openly available data, code, and infrastructure alone does not guarantee accountability to the public or prevent misuse of AI models and, relatedly, widely available model weights are not inherently dangerous. Open resources must be coupled with policy interventions and policymakers should be clear-eyed about what policies will be needed to maximize these benefits and the feasibility of enforcing them.

Below, we delineate some of the goals supported by access to openly available data, code, and infrastructure and describe policy choices that do or do not support them.[12]

---

[11] AI Policy and Governance Working Group. Comment of the AI Policy and Governance Working Group on the NTIA AI Accountability Policy Request for Comment, Institute for Advanced Study, June 12, 2023; AI Policy and Governance Working Group. Recommendations on Global AI Governance to the United Nations Secretary-General's Envoy on Technology, Institute for Advanced Study, September 30, 2023.

[12] See also, Stanford Human-Centered Artificial Intelligence and Princeton Center for Information Technology Policy, Comment on AI Accountability Policy to the National Telecommunications and Information Administration, June 12, 2023.

Advancing Innovation

Making foundation models more widely accessible, with appropriate safeguards, could drive innovation in research and business–capitalizing on the promise of public benefit. Study and use of state-of-the-art AI models, including Large Language Models and other models like AlphaFold, may lead to improvements in performance, safety, and scientific breakthroughs across various domains. These potential benefits can best be realized if other AI model assets, such as model training data, are also made widely available, and if models are not subject to restrictive licenses. Areas that stand to potentially gain from a commitment of ensuring the wide availability of AI tools and systems include, but are not limited to, innovation and novel applications in public health, biomedical research, and climate science that might be scaled in the public interest. Any decision to constrain the availability of dual-use open foundation models must carefully weigh and consider these potential societal and economic benefits.

Reducing the Concentration of Expertise and Power

Today, a significant portion of the resources required to develop the most advanced closed foundation models, including data, compute, and expertise, are held by a few leading companies and organizations. This imbalance in resources and expertise raises concerns about the potential for disproportionate control over these critical AI systems. Promoting wider access to safe, trustworthy, and accountable open foundation models can help address these concerns and foster a broader and more dynamic ecosystem. However, gatekeeping access to model weights alone will not necessarily reduce concentration of power and talent, as compute resources, data, and expertise are also key factors. Fulfilling the Biden-Harris Administration's commitments to market competition and fostering ingenuity will require the protection of pathways to grow and access an ecosystem of open foundation models.

Enabling Transparency as a Tool for Accountability and Public Understanding

Transparency is a valuable by-product of open systems, and when paired with other policy levers such as documentation and disclosure, it can be a powerful tool for accountability. However, transparency alone does not guarantee accountability, and further safeguards and incentives are needed to ensure a broad cross-section of actors are able to support the translation of any increase in the availability of information about AI models.[13] [14] To create a robust ecosystem of accountability, policymakers should

[13] Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in practice*, *17*(4-5), 663-671.

[14] Davies, T., Walker, S., Rubinstein, M., & Perini, F. (Eds.). (2019). *The State of Open Data: Histories and Horizons. Cape Town and Ottawa: African Minds and International Development Research Centre*. Chattapadhyay, Sumandro and Davies, Tim. Chapt. 12: Land Ownership: Open Data and Land Ownership.

incentivize the growth of a diverse third-party evaluation sector (e.g., independent auditing) through targeted research funding, requirements that government-procured AI systems be subject to such evaluation, or other policy levers. Fostering public understanding of both open and closed AI systems is crucial for building societal resilience to the complex challenges posed by new technologies, and as initiatives like the Open Government Partnership have demonstrated, may also have the potential to strengthen democratic governance.

Supporting Safety and Security

Open foundation models can support U.S. safety and security aims, especially if combined with a spectrum of access process (described below) that could enable more people to probe AI systems and potentially identify risks. Open foundation models may also enable two types of risks: identified and emergent. Identified risks are specific, well-defined risks that have been recognized and studied, if not mitigated, such as bias, information integrity, child sexual abuse material, cybersecurity, privacy concerns, and misuse in sensitive domains including biotechnology, chemical, biological, radiological and nuclear defense (CBRN). Emergent risks may arise from the complex, poorly understood, and evolving capabilities of advanced AI systems and, therefore, may be challenging to mitigate. This category includes risks related to the potential development of dangerous capabilities, as well as the emergence of agents (i.e., AI systems capable of performing tasks autonomously). Providing wide access to open foundation models can make it more difficult to prevent and mitigate both identified and emergent risks because, once a model is released, developers do not have the option of revoking, restricting, or monitoring deployer or user access when a new risk is identified as may theoretically be possible with closed foundation models.

Threshold Gradients, Not Binaries

To advance the goals established above, policymakers should consider the circumstances under which heightened scrutiny of an open foundation model may be warranted. The appropriate “thresholds”–that is, benchmarks triggering action–at which models or actors should be subject to additional oversight must be carefully determined and be considered across a spectrum.

In designing thresholds for foundation models, it is crucial to distinguish between (i) the threshold construct (e.g., computational resources, model performance, or societal impact), (ii) the threshold metric that operationalizes the construct (e.g., floating point operations per second or FLOPs, accuracy of specific benchmarks, or the number of downstream applications), (iii) the threshold value which, if exceeded, triggers an action (e.g., 10^26 FLOPs), and (iv) the triggered action itself (e.g., information disclosure).

While the first three components are closely related, the triggered action should be clearly distinguished and aligned with specific governance goals.[15]

The Biden-Harris Administration's use of a compute threshold of 10^26 FLOPs, as outlined in the Executive Order, attempts to capture advanced AI systems pushing the boundaries of AI capabilities. While such a threshold may help delineate certain high capability models, it is important to recognize that compute power alone is not a comprehensive measure of a model's potential risks or societal impact. As the Administration refines its approach to AI governance, it should consider additional factors, such as model performance on specific benchmarks and the extent of a model's integration into society, to develop a more nuanced understanding of which AI systems warrant closer scrutiny and reporting requirements.

A one-size-fits-all approach or a single threshold metric is inadequate for governance because different AI systems and their outputs present unique challenges and risks. For example, Stable Diffusion 2 requires significantly less computational resources to train compared to many prominent large language models; yet its potential risks should not be overlooked due to its lower computational demands. Innovations in model development may further diminish the effectiveness of compute-based thresholds to identify possible risks posed by the most highly capable open foundation models; rather than relying on compute power as the *sole* threshold, a more comprehensive set of factors should be considered.

Evaluations can serve two distinct roles in the context of thresholds for foundation models. They can either be an action triggered by a threshold (e.g., if a model exceeds a certain level of compute power, it must undergo a specific evaluation) or be used as a threshold metric itself (e.g., if an evaluation reveals that a model can perform a high-risk task, further actions may be required). In the latter case, thresholds can help identify, delineate, and filter models, and should be viewed as indicators that may prompt further action rather than definitive risk measures. Developing reliable and widely accepted methods for modeling and quantifying risk is an ongoing area of research that requires significant attention and investment. And as regulators define reporting thresholds for AI models, they must clearly specify the relevant factors of interest, which may include other indicators such as performance benchmarks.

Technical evaluations must be combined with the scrutiny of subject matter experts to develop a comprehensive understanding of an AI system's strengths, weaknesses, and implications across different scenarios. The specific evaluations should be tailored to

[15] Bommasani, R. (2023). Drawing Lines: Tiers for Foundation Models. Stanford Center for Research on Foundation Models.

the risks relevant to a stage and context. However, many technical and conceptual hurdles remain to be overcome to stand up an evaluation sector.

Ultimately, government needs to better articulate its key concerns, both in terms of specific threat scenarios[16] and how AI capabilities might enable or exacerbate those threats. This clarity is necessary to craft targeted governance strategies, including appropriate thresholds.

By developing more detailed threat models, risk assessment thresholds, and governance objectives, governments will be better equipped to provide guidance on essential evaluations at key stages: pre-open sourcing, pre-deployment, and post-deployment. A gradient- or spectrum-based framework for assessing and triggering risk across multiple dimensions is preferable to the binary classifications of open and closed foundation models.

Toward a Spectrum of Access to Open Foundation Models

The NTIA request for comment has framed the debate around the matter of whether or not model weights should be made widely available. However, a proper understanding of the threats and benefits associated with model weights comes from a broader consideration of what is disclosed, how it is disclosed, and who gets access to the model, model weights, or other assets. We recommend that the Biden-Harris Administration consider governance approaches to open foundation models such as forms of staged and structured access, with a policy bias towards openness. Such frameworks would need to clearly specify the types of capabilities, potential harms, or misuse scenarios they aim to address through different governance efforts. The following section explores new access pathways that could help the Administration balance risk mitigation with the promotion of beneficial uses of open foundation models.

Forms of staged released and structured access to open foundation models involve providing controlled access to a model's components while limiting access to its internal information.[17] This approach can be implemented through cloud-based interfaces or platforms, which allow for granular control over who can access the model, for what purposes, and under what conditions. If well designed, staged and structured access approaches can support many of the goals of open access to AI models—especially increasing transparency for accountability and public understanding—while retaining the safety and security benefits of not fully releasing model assets.

[16] Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., ... & Dafoe, A. (2023). Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.
[17] Bucknall, B. S., & Trager, R. F. (2023). Structured Access For Third-Party Research On Frontier Ai Models: Investigating Researchers' Model Access Requirements.

A spectrum of access regime (staged release, structured access, etc.) should mandate that developers share certain information with the entities responsible for testing, red-teaming, and evaluating models. This mandated information sharing is necessary to facilitate effective and comprehensive assessment.[18] The question of what information should be disclosed and to whom is a critical consideration in the governance of AI systems. While transparency is important for building trust, ensuring accountability, and mitigating risk, there may be cases where the full disclosure of certain details could lead to unintended consequences or enable misuse. Developing clear guidelines and protocols for information sharing, including considerations of disclosure to regulators, researchers, and the public, will be an ongoing challenge as the AI landscape continues to evolve.

A structured or staged access regime can help identify vulnerabilities and risks in a controlled manner. To save time and effort, however, it may be desirable for newly discovered flaws or weaknesses to be shared with other model developers, too. The U.S. government can experiment with incentives similar to incident reporting systems in other industries. Just as the aviation sector shares information about failures and vulnerabilities, government can encourage AI companies and developers to share learnings and failures. To facilitate this sharing of learnings and failures, we propose the creation of a Common Vulnerabilities and Exposures (CVE) system for AI, similar to what exists in the cybersecurity industry. This “AI CVE” would serve as a centralized database where AI developers and companies can report and catalog identified vulnerabilities, failures, and potential risks associated with AI systems. This would enable developers of models equivalent to one with a newly discovered vulnerability to assess and mitigate similar risks in their own systems.

A structured or staged access regime will need to balance expert and broad stakeholder involvement, tailored to the nature of the risk. For highly specialized fields with a need for proprietary or classified knowledge–CBRN non-proliferation would be an example–a focused group of domain-specific experts and authorized bodies is essential. This ensures decisions are informed by the deepest available expertise while maintaining safeguards against the dissemination of sensitive information. Conversely, for

[18] Developers should provide specifics on the structured access mechanism itself, including the hosting platform, the capabilities and limitations of the platform, and any restrictions on user actions such as running scripts or fine-tuning models. High-level information about any obfuscated elements should also be shared. Where appropriate, additional relevant information should be provided to enable deeper evaluation and auditing. This may include model checkpoints, user interaction logs, fine-tuning datasets and code, training data, model design parameters, evaluation results, supported input/output modalities, and integrated tool use capabilities.

considerations such as societal impacts and bias, a wider, more diverse array of participants is beneficial, including from academia and civil society.

It is worth noting that piloting staged and structured access programs may be slow, hindering progress in areas where wide access is beneficial. Intellectual property, privacy concerns, unresolved expectations around liability and safe harbors, as well as the allocation of costs pose further challenges. To address these issues, the Biden-Harris Administration may wish to proactively design and implement voluntary spectrum-of-access pilots, focusing on minimizing red tape.

Conclusion

The decision around whether model weights should be made widely available or not requires a full understanding of the threats and benefits associated with open foundation models. A more targeted and nuanced discussion of open versus closed foundation model access, including spectrum of access considerations, is essential.

By proactively designing and implementing staged- and structured-access pilots, policymakers can facilitate progress while addressing potential challenges. In doing so, they should develop clear guidelines for information-sharing among developers, regulators, researchers, and the public. This sharing could be tailored to specific threats and contexts, enabling effective risk assessment and mitigation as well as guidance on essential testing at key stages of model development and deployment. For certain well-defined risks and uncertainties, some precautionary friction may be desirable to ensure appropriate safeguards are in place before widespread access is granted.

The development and deployment of AI models, regardless of their degree of access, should be governed by the principles and practices outlined in initiatives such as the Biden-Harris Administration's Blueprint for an AI Bill of Rights, Executive Order on AI, AI Risk Management Framework, and draft OMB memo on AI in government. Protections and accountability measures developed for open foundation models should aim to harmonize with the broader ecosystem of AI governance, while recognizing the unique considerations that may apply across the spectrum of model openness.

We are grateful for the opportunity to contribute these comments and applaud the NTIA for its leadership in encouraging accountability in the development, deployment, and use of AI systems. As the Biden-Harris Administration navigates the complexities of AI governance, we hope this analysis will provide valuable insights and recommendations, increasing public trust and the responsible adoption of AI in the public interest.

Thank you for your consideration. Please contact aipolicy@ias.edu with any comments or questions.

Sincerely,

AI Policy and Governance Working Group*
Alondra Nelson
Andrew Trask
Ben Garfinkel
Christine Custis
Dan Hendrycks
Deep Ganguli
Dorothy Chou
Helen Toner
Irene Solaiman
Jaan Tallinn
Janet Haven
Marc Aidinoff
Marietje Schaake
Matthew Salganik
Miranda Bogen
Nathan Lambert
Rishi Bommasani
Sébastien Krier
Solon Barocas
Stephanie Ifayemi
Suresh Venkatasubramanian
William Isaac
Zoë Brammer

*Members of the working group are participating in their personal capacities and these recommendations do not reflect the perspective of any of the organizations with which they are affiliated.