SUBMITTED VIA REGULATIONS.GOV

March 27, 2024

Stephanie Weiner
Chief Counsel
National Telecommunications and Information Administration
U.S. Department of Commerce
1401 Constitution Avenue NW, Washington, DC 20230

RE: Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

Dear Ms. Weiner:

Intel Corporation ("Intel") appreciates the opportunity to provide comments to the National Telecommunications and Information Administration ("NTIA" or "Agency") in response to its request for information related to the Agency's responsibilities under the White House Executive Order ("E.O.") on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023. Specifically, the E.O. directs NTIA to conduct a public consultation process and issue a report on 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.

Intel plays an important role in artificial intelligence (AI). Intel's full spectrum of hardware and software platforms offer open and modular solutions which support AI workloads and fuel emerging usages like AI at the edge, pioneering innovations that will advance the future of AI and help to solve the world's most complex challenges. For example, in healthcare and life sciences, we accelerate research and patient outcomes with faster, more secure, and more accurate analysis across precision medicine, medical imaging, lab automation, and more. For manufacturing, we transform data into insights that help companies optimize plant performance, minimize downtime, improve safety, and drive profitability. Intel is committed to advancing AI technology, responsibly and contributing to the development of principles, international standards, best practices methods, tools, and solutions to enable a more responsible, inclusive, and sustainable future.

As policymakers consider approaches to develop frameworks for the responsible development and deployment

of AI technology, it is important to recognize that open development of AI is essential. There is broad concern that due to the nature of advanced model development and its reliance upon significant compute infrastructure, advanced AI technologies have the potential to be controlled by the hands of a few. The opacity and centricity of that possible future could pose risks to democracy and capital systems. Technology and AI innovation are best served through open, neutral, horizontal competition. Open AI development is essential to facilitate innovation and equitable access to AI, as open innovation, open platforms, and horizontal competition help offer choice and build trust. Fostering an open ecosystem for AI that leverages standardized processes and removes barriers to open source tools and solutions can facilitate innovation to thrive and create broader access to AI technology. However, to achieve these benefits, open development may comprehend accountability for responsible design and implementation to help mitigate potential individual and societal harm. This includes establishing robust security protocols and standards to identify, address, and report potential vulnerabilities. Openness not only allows for faster advancement of technology and innovation, but also faster, transparent discovery of potential harms and community remediation and address.

In contemplating this topic, it is important to understand not just the benefits, but also specific risks related to widely available weights, how different levels of availability may impact these risks, and the best approaches to mitigate them. A few examples of potential risks are unintended harm, malicious harm, circumvention or lack of safety controls, and new, unforeseen emergent risks. For an open ecosystem we must find the right balance between the structural advantages and importance of open development, while also considering and addressing potential risks from access to model weights, datasets, and training information, as well as implementing appropriate safety testing and guardrail details for advanced AI models.[1] When evaluating the potential for harm, all data sets are not created equal. Risk management of data should consider not only dataset size, but also classification of data – health and life sciences data may require different safeguards to balance the potential for life improving innovations versus the risks of malicious actors.

We offer the following input on NTIA's efforts related to dual use foundation models with widely available weights as outlined in the Agency's notice.

QUESTIONS

1. *How should NTIA define "open" or "widely available" when thinking about foundation models and model*

[1] *See*, https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf

weights?

A definition of "open" or "widely available" depends on the objectives that NTIA aims to promote. If an NTIA considers the question of "open" and "widely available" to address limitation and promote broader access to frontier or foundation models, it may be useful to look to the experience of the software ecosystem. The Open Source Initiative (OSI)[2] has developed a definition that requires open, permissive, and generally unrestricted access to and use of software to be considered "open". In this regard, it is promoting free, broad access to an ever-expanding corpus of source code to promote quality and innovation in software development. Others in the ecosystem promote their software as "open" if it is merely available for download, albeit with significant restrictions on access and/or use. Each has their attendant benefits and risks. While the OSI model has indeed improved the quality of software and promoted innovation, it also makes it easier to develop applications used to support criminal activity. Conversely, some have "open-washed" their software by promoting use-restricted software as "open" only to create software lock-in to a broader product portfolio.

In the context of AI solutions, a more nuanced approach to defining "open" and "widely available" may be in order. NTIA may consider a definitional approach that promotes an open, permissive access to enabling AI technologies (data, models, etc.,) to promote broader innovation in the AI ecosystem, while still allowing for the use of certain legal and technological restrictions on access or use of AI models to address misuse or abuse in AI applications. That is, a definition of "open" that promotes open access to and permissive use of certain levels of model architecture and model weight information, and responsibility for how and where those powerful models are used and their ultimate impact on society.

LFAI & Data[3], an umbrella foundation of the Linux Foundation that supports open source innovation in AI and data, promotes a Model Openness Framework[4], which is a ranked classification system that rates AI models based on their completeness and openness, from open model, open tooling, to open science. Model openness encompasses more than model weights and includes but is not limited to datasets, topology, training and inference code, pre-processing code, hyperparameters, and evaluation metrics. A question to consider is how easily a third party can replicate advanced AI models as all parameters listed above, in

[2] https://opensource.org/
[3] https://lfaidata.foundation/
[4] https://docs.google.com/document/d/1RUNrs4flAsYsikXTPu1jWBH1BAumCyeG/edit?tab=t.0#heading=h.gjdgxs

addition to model weights, need to be understood.

c. *Should "wide availability" of model weights be defined by level of distribution? If so, at what level of distribution (e.g., 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be "widely available"? If not, how should NTIA define "wide availability?"*

Level of distribution should not be used to define "wide availability" of model weights. Rather, NTIA should focus on components related to risk assessment. The LFAI & Data Model Openness Framework[5] can be a helpful reference to define "wide availability" of model weights. Also, refer to the response to question #1 above.

d. *Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access?*

Release of foundation models present a gradient of openness options.[6] A few examples are listed below.

1) Hosting the model and allowing access only through an API enables the host to monitor all requests, scrutinize requests, and maintain safeguards.
2) Enabling others to host the model and/or ensure that requests and safeguards are controlled. In this case misuse is likely not observed, as prompts are not monitored and insight into model use is limited.
3) Releasing model code or details to "vetted" users can enable watermarking/fingerprinting for tracking the proliferation of models. This scenario assumes control is only implemented by user selection/vetting.
4) Releasing model code to open source so that anyone can have access.

As model providers start at option 1 and transition to options 3 or 4, this approach can assist to initially discover issues and capabilities of the model. Additionally, while lower degrees of openness (e.g., controlled access via APIs) may offer lower potential for downstream innovation and scrutiny than an open source, licensed release of foundation models, lower degrees (i.e., gradients labelled 1, 2, and 3 above) can help mitigate the risk of misuse or abuse of the model.

[5] *Id.*
[6] *See,* https://arxiv.org/abs/2302.04844

i. *Are there promising prospective forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?*

Yes, the security of AI solutions, including securing the release of AI models, is an area of active research and development within academia, industry, and industry consortia. As a result, the ability to secure models once they are released into the ecosystem – even under generally open and permissive license models (if not pure open source licenses) – will continue to improve. Between fully closed and fully open modes of access, there generally exists some form of controlled access (e.g., staged release, selected access by "trusted" parties, API-based) to some elements of the model, including barring access to others. While limited forms of access can still provide value beyond that of fully open models (e.g., staged releases for incremental model evaluation or API access to fine tune and build downstream systems) it is likely that these will fall short of the benefits of full open sourcing.

2. *How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?*

When assessing the risks associated with the release of a model, the model architecture and model weights need to be considered together. Assuming that the question presumes that a model architecture is widely available, the risk of abuse or misuse of the model may be mitigated, at least for some time, by limiting the open availability of the full set of model weights. Informally, publicly available model weights make it easier to modify the model, reduce in-built guardrails, and attempt to leverage them for propagating various societal harms and malicious uses. Even when the model is being fine-tuned for non-malicious use, unless proper model testing and evaluation practices are commissioned, the model may have unintended consequences leading to harmful societal impact.

One can also use non-public models (e.g., API-based models) for propagating harms to some extent. However, these can be better controlled (based on API usage) and proper interventions can be put in place. Closed models may also be susceptible to malicious use, given that current safeguards may not be robust enough to withstand adversarial attacks.

See the chart below comparing risks of widely available and non-public model weights. NTIA could consider defining formal tiers of open software. Tier 1 might allow access to model weights only, while Tier 2 could also permit data sources. Tier 3 could open-source the source code used for model training, and Tier

4 might add any pre- and post-processing steps, or details like reinforcement learning from human feedback (RLHF)[7]. That way, lower tiers would allow scrutiny and use of the models while upper tiers would enable full reproducibility. Such tiering systems are common in other safety metrics (e.g., L1-L5 defined for advanced driver assistance systems (ADAS). This approach would also provide a mechanism for consistent "open access", as currently, entities label products as open source when they are actually open models, thereby creating false expectations in the ecosystem.

Risks	Open Weight Foundation Models	Non-public Model Weights
Policy enforcement based on various Risk Management Frameworks	Difficult	Easy
Decommissioning of older models	Difficult	Easy
Enforcing critical updates to the model	Difficult	Easy
Controlling Model parameters for various risk levels and thresholds	Difficult	Easy
Unacceptable use Risk Controls	Difficult	Moderate
Unforeseen Risk Control	Difficult	Moderate
Model Transparency	Easy	Moderate
Model Testing, Red Teaming	Easy	Difficult
Proliferation of Technologies/IP	Widespread	Controlled
Marginal Risk of open vs closed/pre-existing technologies	Limited (e.g., Misinformation, Deepfakes)	-
Risks associated with downstream applications that use Foundation Models	Similar	Similar

c. *What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?*

There are two types of risks to highlight for NTIA's consideration. One set of risks is the intentional use of an AI model to cause harm or violate human rights and national security. Avoiding open release of frontier models or foundation models of this capability/fidelity under an open architecture or open weights model could address this risk. That said, once these models are released (even proprietarily through

[7] *See*, https://huggingface.co/blog/rlhf

commercialization), it may not take long for variants of it to emerge in open source across the world. These may approximate the same capability for a narrower set of use cases. This is because the fundamental architecture of almost all foundation models today is based on a small set of ideas – transformers, diffusion models, contrastive learning, etc., and there is little intellectual barrier to entry. As such, policymakers should prioritize identifying and mitigating malicious use of the technology rather than prescriptive regulation of the technology itself.

The other type of risk is an AI system which is empowered to perform consequential tasks, yet it operates in a manner that deviates from its intended design and causes harm or violates laws and principles in an unintended manner. The risks due to uncertainty in a model and/or lack of understanding of edge-scenario behavior of the model is a much higher risk that can be exacerbated by adversarial attacks on the model. Adversarial AI influences or compromises a commissioned AI system to cause it to perform in an unintended manner. Addressing this risk should be done through increased transparency of datasets and pretraining/finetuning processes. For adversarial use, widely available weights and training data enable a faster path for malicious activity – especially if all aspects are being shared, thereby permitting complete reproducibility of its results. Explaining the principles that lead to a leap in capabilities compared with previous models can sometimes generate more effective reproduction than just providing weights and data. For example, the principle of reinforcement learning from human feedback can facilitate major improvement of results even if none of the weights and specific data are shared.

Assessing risks associated with a model's reproducibility are linked to the foreseeable risks of the model, itself. The availability of model weights alone is hardly sufficient to reproduce the model, however, the availability of architecture, hyperparameters, datasets, topology, training and inference code, evaluation metrics, and other detailed information can aid in reproducing the model. As openness to the public increases, the more others can leverage it to reproduce, enhance, and adapt the model to other domains, potentially adding risks. However, the ability to build a performance model relies on compute availability and data parallel training. This is necessary for a malicious actor even if the full algorithm was published. Yet, making the algorithm public could also enable mitigation plans and help balance the risk.

d. *Could open foundation models reduce equity in rights and safety-impacting AI systems (e.g., healthcare, education, criminal justice, housing, online platforms, etc.)?*

Foundation models can have negative impacts by being misused or by being intrinsically biased, unsecured,

or inaccurate. However, open-ness or closed-ness of a foundation model is not likely directly related to an increase or decrease in equity. API calls to a closed model can be just as vulnerable to biases as compared to directly using an open model. This is a fundamental issue for all machine learning models that are "blackbox" in nature. In the case of unintended consequences in a consequential deployment, frontier models cannot be validated across all aspects of functionality similar to other software products. Open models (datasets, architecture, weights, training scripts, testing performed) can help developers to improve safety issues and reduce the barrier for malicious actors to circumvent safety guardrails. Even transparency without full reproducibility can lower the barrier for model attacks and compromised safety. Transparency in testing and intended use cases can assist downstream users to select appropriate models and understand risks to adapt specific models to specific use cases. In an open ecosystem, harm may result from malicious actors or well-meaning developers that break safety guardrails unintentionally. However, the latter is easier to mitigate with standard safety benchmarks.

e. What, if any, risks related to privacy could result from the wide availability of model weights?

Open weight models present increased marginal risks related to deepfakes as they substantially lower the barrier to generating such content.[8] Safeguards for closed models, including watermarking, are relatively more effective in this area, and monitoring closed models can deter generating such imagery, especially of real people. Voice cloning based scams may also pose an increased potential threat with open models compared to closed models (although there isn't much documented evidence of these harms).

3. What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?

c. *What benefits do open model weights offer for competition and innovation, both in the AI marketplace and in other areas of the economy? In what ways can open dual-use foundation models enable or enhance scientific research, as well as education/training in computer science and related* fields?

Open model weights have the potential for spurring further innovation in AI. For example, almost all innovation in AI to-date has been due to openly available infrastructure[9]. Open model weights are likely going to aid researchers to find impactful and beneficial use cases of AI that will be overlooked by narrow

[8] https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf
[9] Openly available infrastructure encompasses more than model weights, including but not limited to architecture and dataset transparency.

and immediate commercial interests of proprietary model vendors. An example of this is applying leading-edge AI principles to open scientific problems. Open model weights also spur more startups and innovations, enabling startups to quickly prototype without access to immense capital, fostering a more competitive landscape.

d. How can making model weights widely available improve the safety, security, and trustworthiness of AI and the robustness of public preparedness against potential AI risks?

Open model weights will allow researchers at the intersection of AI and safety, security, trust, and ethics to research methods to further improvements in model fairness and trustworthiness. This kind of research is essential to understand data strategies, model architectures, and how to learn from human feedback so that models can be made more trustworthy and robust.

e. Could open model weights, and in particular the ability to retrain models, help advance equity in rights and safety-impacting AI systems (e.g., healthcare, education, criminal justice, housing, online platforms etc.)?

Yes, this approach presents an opportunity that initial deficiencies in models (e.g., racial or gender biases) can be studied openly by researchers and techniques developed can debias these models. A research paper by Intel Labs demonstrates how to mitigate social biases in vision-language foundation models.[10] Such work will not be possible without access to the original model weights and is essential to ensure that foundation models can fairly and equitably represent all sections of populations. The ability to retrain these models to address safety vulnerabilities and bias, especially given the specificity of downstream applications and deployment contexts and their associated risks, would be key to improve the equity and safety of such models.

4. *Are there other relevant components of open foundation models that, if simultaneously widely available, would change the risks or benefits presented by widely available model weights? If so, please list them and explain their impact.*

Please see the response to question #2.a. above.

5. *What are the safety-related or broader technical issues involved in managing risks and amplifying benefits*

[10] https://arxiv.org/abs/2312.00825

of dual-use foundation models with widely available model weights?

a. What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?

Amongst other areas of work, capability evaluations are being investigated by the recently created U.S. Artificial Intelligence Safety Institute and the supporting AI Safety Institute Consortium. In parallel, a number of processes and techniques, such as the use of responsible AI councils to evaluate AI models, red-teaming, penetration testing, model cards, dataset evaluations, security reviews, etc. may be employed to help identify and mitigate risks associated with model development and the release of such models. Additionally, for an open framework, benchmarking and adversarial testing can provide information that a model is capable of inflicting serious harms without added guardrails (i.e., catastrophic events), and thus limiting the release of the architecture and weights to "known good actors" would be important.

c. What are the prospects for developing effective safeguards in the future?

While prospects are promising, this is an extraordinarily difficult problem to solve not unlike what is seen today in the hacker/cybersecurity space. It will require constant evolution of the technology.

f. Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model? To the extent possible, please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.

Model weights, metadata, model architecture, datasets, training/inference/evaluation codes, hyperparameters, data processing code, and evaluation metrics are components of a foundation that would be needed to analyze, evaluate, certify, or red-team the model; applicable research papers can also assist in this effort.

g. Are there means by which to test or verify model weights? What methodology or methodologies exist to audit model weights and/or foundation models?

HELM[11] is a great resource to evaluate foundation models; it encompasses model accuracy, performance, bias, toxicity, and other means for 80+ different scenarios.

7. What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the

[11] https://crfm.stanford.edu/helm/classic/latest/

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?

d. *What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights?*

The U.S. government should take an active role in setting metrics for risk and creating standards for best practices for the responsible development of deployment of foundation models. Leveraging the various initiatives by the National Institute of Standards and Technology (NIST), other multilateral organizations, and relevant international standards bodies, the U.S. can model for other countries the importance of prioritizing best practices and international standards development for foundation models.

8. *In the face of continually changing technology, and given unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future?*

b. *Noting that E.O. 14110 grants the Secretary of Commerce the capacity to adapt the threshold, is the amount of computational resources required to build a model, such as the cutoff of 10^{26} integer or floating-point operations used in the Executive order, a useful metric for thresholds to mitigate risk in the long-term, particularly for risks associated with wide availability of model weights?*

The computational resource threshold outlined in Executive Order 14410 is not the best mechanism for establishing thresholds to mitigate risk associated with wide availability of model weights for dual use foundation models. While this approach correlates decently with model capability and measurement, its use was intended to be temporary and not future proof for long-term implementation. The Department of Commerce can consider facilitating this effort through NIST, as it is currently evaluating capabilities and metrics for dual use foundation models through its AI Safety Institute Consortium and other related activities.

***

Intel appreciates NTIA's consideration and welcomes the opportunity to discuss our feedback.

Respectfully submitted,

Angel Preston
Policy Director, Artificial Intelligence