Response to NTIA Request for Comments on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

27 March 2024

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

Subject: Openness in AI Request for Comment

Via Regulations.gov, docket number NTIA–2023–0009

To Ms. Weiner,

Thank you for the NTIA Openness in AI Request for Comment (RFC) released February 2024. We offer the following submission for your consideration.

We are researchers affiliated with The University of California, Berkeley, with expertise related to AI development, safety, security, policy, and ethics. We previously submitted responses to NTIA last year on AI Accountability (https://medium.com/cltc-bulletin/response-to-ntia-request-for-comments-on-ai-accountability-policy-c2aca7e4285c), and to NIST several times over the past three years at various stages of NIST's development of the AI Risk Management Framework (AI RMF).

The debate about "open or closed" foundation models has become contentious in policy and technical communities, but there are middle ground approaches that can help to balance the benefits of openness with the risks from the proliferation of unsecured dual-use foundation models. We emphasize these approaches in our response. A recent survey of more than 1,000 Americans found similarly nuanced considerations, including interest in improving independent researcher access while recognizing risks of open sourcing powerful AI models.[1]

[1] A March 2024 poll conducted by the AI Policy Institute with an online sample of 1014 respondents found that while 71% of respondents believe that academic researchers should be given access to AI models, 74% believe that open sourcing powerful AI models so it's easier for more developers to use and alter powerful AI models without restrictions is a *bad idea*. And 77% believe models should be prevented from producing what many consider to be inappropriate or offensive content. "AIPI March OpenAI Poll Toplines", https://drive.google.com/file/d/1tGgq_3qfhHSPBH_-cnQ7MZcjD7KmysQz/view.

NTIA has used the term “open foundation models" as a shorthand for the more specific term used in the White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, “Dual-Use Foundation Models with Widely Available Model Weights.” However, we note the term “open” is often conflated with “open-source” and can overstate the true transparency of a release strategy while creating a sense of a false binary. In practice there is a spectrum of release methodologies between “open” and “closed”.[2] Throughout our response, we typically use the terms “**unsecured**” and “**secured**” instead of “open” and “closed” respectively, to refer specifically to the question of whether model weights have been made widely available (e.g. downloadable on a public repository).[3, 4]

Here are some of our key comments and recommendations on the NTIA Openness in Artificial Intelligence Models RFC:

- AI models with widely available weights, or unsecured models, can provide important benefits such as enhanced privacy for intended model users, easier auditability, as well as a more widely accessible research and innovation ecosystem.
- However, unsecured models also pose risks, such as various forms of malicious misuse resulting in harm, including to people’s rights and wellbeing and to the safety of the general public. Although both closed and open models can pose some such risks, unsecured models pose unique risks in that safety and ethical safeguards that were implemented by developers can be removed relatively easily from models with widely available weights (e.g., via fine tuning).[5] Knowing that, the developers of some open models do not attempt to implement safeguards in the first place. It is also impossible to ensure that critical security updates or other updates are effectively propagated to all instances of an open model.
    - **Available evidence has demonstrated harmful impacts from misuse of unsecured models**, in particular for child sexual abuse material (CSAM), non-consensual intimate imagery (NCII), disinformation at scale, facilitating cyberattacks, enabling online radicalization, and promoting harmful stereotypes and violence.[6] These risks are likely to continue to disproportionately harm

[2] See, e.g. “The Gradient of Generative AI Release: Methods and Considerations,” https://arxiv.org/abs/2302.04844.

[3] “How to Regulate Unsecured “Open-Source” AI: No Exemptions,” https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.

[4] However, we do not mean to imply that “secured” foundation model developers have implemented perfect information security that would prevent all well-resourced adversaries from stealing weights of their models. See, e.g., Nevo et al. (2023), “Securing Artificial Intelligence Model Weights: Interim Report”, https://www.rand.org/pubs/working_papers/WRA2849-1.html.

[5] See, e.g., “BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B”, https://arxiv.org/abs/2311.00117; “Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation”, https://arxiv.org/abs/2310.06987; and “MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots”, https://arxiv.org/abs/2307.08715.

[6] See, e.g., “Investigation Finds AI Image Generation Models Trained on Child Abuse,” https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse; “The Dark Side of Open Source AI Image Generators,” https://www.wired.com/story/dark-side-open-source-ai-image-generators/; "‘RedPilled AI’: A New Weapon for Online Radicalisation on 4chan," https://gnet-research.org/2023/06/07/redpilled-ai-a-new-weapon-for-online-radicalisation-on-4chan/;

women, minority groups, and vulnerable communities including children and the elderly.[7]

    - Some have argued that the marginal risk of open foundation models is relatively minimal in at least some cases[8]. The notion of "marginal risk" can be a helpful framing to ground the discussion of risks in a broader context. However, some of the most well-documented risks, such as the promotion of harmful stereotypes and violence, were not included in that particular study. We also expect the marginal risks of unsecured model release to grow, especially for the largest-scale and most broadly capable models – including dual-use foundation models as defined in Executive Order 14110 – that could eventually pose the greatest risks of enabling severe harms via malicious misuse. Capabilities of the largest-scale and most broadly capable "frontier" models have tended to increase with larger models and with larger quantities of data and compute used in model training, and we expect continued increases in each of those dimensions for frontier models. Thus, we expect increases in frontier-model capabilities (e.g., writing software code) that also would increase malicious-misuse hazards (e.g., writing malware).
- In addition to investing in upstream protections to prevent a range of misuses of unsecured foundation models, we should invest in downstream protections to prevent specific misuses. However, **we should not rely only on downstream protections.**
    - For example, many advocate requiring U.S. mail-order gene synthesis labs to screen orders as a downstream protection to prevent the creation of bioweapon agents following malicious misuse of an unsecured model. However, it has long been recognized there are mail-order gene synthesis labs in China or elsewhere outside the United States that are much less likely than U.S. labs to follow standards for screening gene synthesis orders, which limits the value of mandatory screening of gene synthesis orders in the United States.**[9]**
    - Model developers are often the sole decision makers when it comes to choices about data, model design, evaluation, and mitigations. Developers of foundation models are also often highly resourced companies or organizations. Requirements for reasonable upstream protections provide accountability for the organizations in AI value chains that have the greatest power to reduce risks to the public from AI systems.

"WormGPT: when GenAI also serves malicious actors," https://eviden.com/insights/blogs/wormgpt-when-genai-also-serves-malicious-actors/; "Counter Cloud - AI powered disinformation experiment," https://countercloud.io/?page_id=307; "How Does Access Impact Risk? Assessing AI Foundation Model Risk Along a Gradient of Access," https://securityandtechnology.org/virtual-library/reports/how-does-access-impact-risk-assessing-ai-foundation-model-risk-along-a-gradient-of-access/).

[7] "Technology-facilitated gender-based violence in an era of generative AI," https://unesdoc.unesco.org/ark:/48223/pf0000387483.

[8] See, e.g., the marginal-risk framework of Kapoor et al. 2024, "On the Societal Impact of Open Foundation Models", https://crfm.stanford.edu/open-fms/paper.pdf.

[9] See, e.g., Tucker 2010, "Double-Edged DNA: Preventing the Misuse of Gene Synthesis", https://issues.org/tucker-2/; Lovett 2024, "DNA bought online could be used to create dangerous pathogens, experts warn", https://www.telegraph.co.uk/global-health/science-and-disease/dna-could-be-bought-online-to-create-dangerous-pathogens/.

- As part of managing the risks of unsecured model release without preventing the benefits, we recommend: **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 secured 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.[10] **The largest-scale or most capable models (including dual-use foundation models as defined in Executive Order 14110) 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.[11] Structured access such as through APIs typically provides more opportunities than downloadable model weights allow, for mitigations such as content filters to prevent misuse, monitoring of usage to identify misuse, and system shutdown or rollbacks if problematic usage is identified. Foundation model developers that publicly release the model parameter weights for their foundation 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. After model weights have been downloaded, the downloading actor can in turn make the model weights available to others, including via distribution channels that can be hard to monitor and harder or impossible to shut down. (See our response to RFC question 8a for additional details on these recommendations.)
- We also recommend openness mechanisms that do not require making a model's parameter weights downloadable. Many of the benefits of open-source systems, such as review and evaluation from a broader set of stakeholders, can be supported through transparency, engagement, and other openness mechanisms besides releasing model parameter weights. For example, **independent researchers should be able to test and audit secured models more readily and with the protections of a legal safe harbor.**[12] Other benefits of open source, such as making AI technologies more widely available, can also be achieved by providing less-resourced actors with cost-free or cost-subsidized access to secured foundation models. In addition to removing monetary barriers, freely-accessible secured models can often be more inclusive to less technologically sophisticated users. The availability of open-source models does not

[10] To make this recommendation less burdensome to implement, we suggest resource-constrained foundation model developers such as academics make use of third-party model hosting infrastructure services (e.g. Amazon SageMaker), especially for models far below dual-use foundation model thresholds. We suggest that infrastructure service providers, in turn, continue to develop their services to take more labor-intensive aspects of secure model hosting, e.g. abuse monitoring, off of the shoulders of model developers.

[11] We draw this recommendation from our 2023 AI Risk-Management Standards Profile for General-Purpose AI Systems and Foundation Models, https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf.

[12] "A Safe Harbor for Independent AI Evaluation," https://sites.mit.edu/ai-safe-harbor/.

solve the many structural barriers, such as access to computing power and data, that many people, communities, and organizations face in participating in AI development.[13] (See our response to RFC question 8a for additional details on these recommendations.)

In the following sections, we provide more detail and additional comments.

Views we express here in this piece are our own, and do not necessarily reflect the views of our employers or funders.

Thank you again for the opportunity to comment on the NTIA Openness in AI RFC. 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 NTIA as you and others consider how best to approach the costs and benefits of AI openness.

Our best,

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

David Evan Harris
Chancellor's Public Scholar, UC Berkeley
Continuing Professional Faculty, Haas School of Business
Senior Research Fellow, International Computer Science Institute, UC Berkeley

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

Nicole Lemke
Visiting Scholar, Center for Long-Term Cybersecurity, UC Berkeley

Evan R. Murphy
Non-Resident Research Fellow, 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

[13] "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI," https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807.

Co-Director, AI Policy Hub, UC Berkeley
Assoc. Research Professor, Goldman School of Public Policy, UC Berkeley

Andrew W. Reddie, Ph.D.
Assoc. Research Professor, Goldman School of Public Policy, UC Berkeley
Founder, Berkeley Risk and Security Laboratory

Stuart Russell, Ph.D.
Distinguished Professor of Computer Science, UC Berkeley
Director, Center for Human-Compatible AI

Bin Yu, Ph.D.
Distinguished Professor of Statistics, EECS, and Center for Computational Biology, UC Berkeley

Our comments on specific items in the NTIA Openness in AI RFC

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

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

b. Is it possible to generally estimate the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks? How do you expect that timeframe to change? Based on what variables? How do you expect those variables to change in the coming months and years?

In a review of data on model trends since 2020 across language, multimodal, biological, and vision, there is the appearance of a trend of openly available models of a particular size and capability level following about a year or two after otherwise similar AI models become available on a closed basis.[14]

There are many methods for comparing model capabilities besides parameter counts and the overall size of the dataset. This initial analysis does not show a universally applicable trend for when a model of a certain size or capability will be released via one mechanism or another, when considering all types of models and all time periods. Prior to 2020, many notable models were open-source and primarily academic examples. However, as models have grown significantly in size, the prohibitive cost of training large models may result in fewer open models

[14] Epoch (2023), "Key Trends and Figures in Machine Learning", https://epochai.org/trends.

of similar capability. This is further complicated by the fact that models are built on top of and in conjunction with one another, particularly ones built for specialized purposes, making the distinction between open and closed blurry. This comparison also does not take into account when a very large model is published that performs well on a variety of tasks and then is fine-tuned or refined in some way.

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?”

For purposes of reducing the potential for malicious misuse of a model, it seems more important to consider whether a model is openly available at all than to consider whether it has been known to have been downloaded 10 times, 10,000 times, or 1,000,000 times. After model weights have been downloaded, the downloading actor can in turn make the model weights available to others, including via distribution channels that can be hard to monitor and harder or impossible to shut down. Furthermore, such secondary distribution cannot be accurately and precisely measured when attempting to estimate the “level of distribution” of a model.

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?

It is not entirely clear to us whether NTIA intends this RFC question 1.d. to ask about I) how a foundation model developer would provide access to download a model's weights etc., or II) how somebody else would use that model after they download it, e.g. to build a customer service chatbot as part of their air travel website. Regarding I), a truly open foundation model would not be hosted on a web application or API, but instead would allow for local downloading and interaction. Regarding II), please see our table below.

There is a spectrum of release methodologies between the binaries of “open” and “closed”.[15] The following table outlines some of the benefits and negative risks associated with each of the several options for how a user would interact with a model. This analysis is not unique to AI applications but mirrors the considerations for software security writ large. Each approach has security benefits and drawbacks associated with it. These risk and cost drawbacks are largely dependent on other details of the system or application enabling access. The overall security infrastructure, training data sensitivity, and prompt information must be assessed to make an appropriate decision on model access and hosting requirements.

[15] “The Gradient of Generative AI Release: Methods and Considerations,” https://arxiv.org/abs/2302.04844.

Interface for hosting or accessing a foundation model (or applications built on the foundation model)	**Benefits**	**Risks and Costs**
Web applications	-Widely accessible yet centrally controlled -Can employ authentication and access controls -Regularly updated through patches -Use monitoring and logging to monitor usage	-Susceptible to web application-specific attacks[16] -If login credentials are not present or correctly managed these models become widely accessible and unattributable to specific users -Relies on one developer to update and maintain security standards instead of the larger community
API	-Widely accessible yet centrally controlled -Access control and authorization -Rate limiting to control the number of requests -Regularly updated through versioning -Use monitoring and logging to monitor usage	-Susceptible to API-specific attacks[17] -If login credentials are not present or correctly managed these models become widely accessible and unattributable to specific users -Relies on one developer to update and maintain security standards instead of the larger community
Local Hosting	-Accessibility limited -Enhanced control over the data (privacy, security) -Eliminates cloud-based hosting risks -Can be internet-gapped, segmented, or isolated allowing for fewer instances of unauthorized access and lateral attacks -Minimized attack surface	-Unmonitored by upstream developers or cloud providers -Requires independent security measures specific to the AI application be applied in addition to traditional best network security practices -Enhanced backup and recovery processes needed -Additional independent data security practices[18]
Edge Deployment	-Widely accessible and distributed control -Enhanced resilience -Minimized attack surface per	-Edge devices typically have more variability which could lead to misconfiguration or difficulty ensuring proper

[16] https://owasp.org/www-project-top-ten/.

[17] https://owasp.org/API-Security/editions/2023/en/0x11-t10/.

[18] Particularly in regards to GDPR and the EU AI Act.

	device -Enhanced control over the data (privacy, security) as sensitive and secure data is processed only at certain nodes	security controls on every device, patch management, etc. -Overall expands attack surface by introducing more interconnected devices

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

As we discuss in some depth in our response to RFC Question 8a, drawing upon our AI Risk-Management Standards Profile for General-Purpose AI Systems and Foundation Models[19] and our related policy brief[20], we recommend the following:
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 secured or structured access release such as through APIs, and 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 also recommend freely accessible secured foundation models as another approach that strikes a favorable benefit-risk balance. When models are available cost-free (or cost-subsidized), but still provided via structured access such as behind APIs and hosted UIs, they could achieve many of the potential benefits of widely available model weights that NTIA noted in this RFC, such as fostering growth among less resourced actors, and helping to widely share access to AI's benefits. Moreover, they could achieve those benefits without giving up the risk-mitigation benefits of structured access, such as monitoring server traffic and logs for indicators of abuse and other important risks. We elaborate on this approach in our response to NTIA RFC Question 8a as well.

There are many other forms of access and transparency that can further help strike a more favorable benefit-risk balance. One is to provide greater access to the training data for broad review. Another is to provide independent researchers with full model access, and to provide a legal safe harbor for good faith independent academic evaluation (https://sites.mit.edu/ai-safe-harbor/). Different forms or modes of access will be more or less meaningful for different

[19] AI Risk-Management Standards Profile for General-Purpose AI Systems and Foundation Models, https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf.

[20] Policy Brief on AI Risk Management Standards for General-Purpose AI Systems (GPAIS) and Foundation Models, https://cltc.berkeley.edu/publication/policy-brief-on-ai-risk-management-standards-for-general-purpose-ai-systems-gpais-and-foundation-models/.

audiences and it is important to note that widely available model weights only represent a particular type of access that promotes adaptation of models and is most useful for developers. Accessibility for other groups and communities may be better supported through other means, for example through detailed documentation of training, capabilities, and limitations in the form of model cards or other such documentation methods (see, e.g., “Model Cards for Model Reporting,” https://arxiv.org/abs/1810.03993).

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

a. 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?

Openly releasing model weights is irreversible and means that AI developers no longer have a mechanism to restrict access to people who are using the AI model to cause harm. It is therefore critical to fully consider the risks associated with foundation models with widely available model weights prior to release. The AI sector at large has no single, standardized approach towards risk management for foundation models, and there is insufficient transparency around (1) how risks are assessed and (2) how thresholds are determined. However, many industry and academic organizations have their own frameworks and policies that they act upon individually.

Due to this fragmented risk management ecosystem, decisions appear to be made in an ad hoc fashion and often appear to stem from competitive corporate dynamics. This is deeply troubling given the scale and scope of risks associated with foundation models.

Here we focus on risks to individuals, communities, society, and the environment; in particular, the risks associated with widely available model weights include risks to the public interest, democracy, public safety, and the protection of human rights. These risks include all of the following:

- The promotion of harmful stereotypes, violence, and online radicalization
- The potential manipulation or persuasion of large groups of individuals to take harmful actions, via the proliferation of chatbots without adequate safeguards
- The spread of non-consensual intimate imagery (NCII) including child sexual abuse material (CSAM)
- The spread of other synthetic media including audio, images, or video, with the intention to deceive voters or otherwise disrupt democratic processes
- Increased ease through automation of multidomain threats relevant to national and international security (e.g., terrorism, cyber-attacks, environmental degradation, and economic instability)

It is important to note that all of these risks also theoretically exist with “closed” models as well, but these model developers have more mitigation options available to them. For some risks, the layers of resistance or mitigations may provide a high enough barrier to dissuade successful adversary action. Many of the frontier model developers have committed to “responsible scaling” policies or related policies which outline the circumstances that would cause them to pause or retract models exhibiting dangerous behaviors.

Additionally, an important issue with risks associated with widely available model weights is the framing of “safety”. Models with widely available weights can provide substantial benefits for increasing *safety for intended users*, in terms of reliability, security and privacy for those intended users. However, models with widely available weights also can pose substantial risks of reducing *safety for the general public* due to misuse by unintended users.

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

Many foundation model developers spend extensive time and resources working to better align their models with desirable human values including equity. For example, prior to the launch of GPT-4 OpenAI spent six months on safety testing, risk assessment, and iteration, including refining the model with reinforcement learning from human feedback (RLHF) to dramatically reduce instances of hate speech, discriminatory language, harassing and demeaning content, and incitements to violence.[21]

Prior to Google DeepMind’s recent open release of the foundation model Gemma, the model was evaluated and tested for representational harms, sexual abuse and exploitation, harassment, violence and gore, hate speech, memorization of training data, as well as chemical, biological, radiological, and nuclear (CBRN) risks.[22] The open release of Gemma was accompanied by Terms of Use, a Prohibited Use Policy, and a Responsible Generative AI Toolkit to help users of the model to apply best practices. These measures are intended to counteract some of the known harms that can come from open releases, which per Gemma’s Model Card may include the perpetuation of biases, generation of harmful content, misuse for malicious purposes, and privacy violations.[23] This documentation is more comprehensive than is typical, but will not in and of itself prevent all expected harms or prevent malicious actors from misusing the model. Google DeepMind has fewer options to prevent misuse of Gemma than it does for its secured Gemini models. Knowing this, Google DeepMind has made Gemma significantly smaller and less capable.

[21] “GPT-4 System Card,” https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[22] “Gemma: Introducing new state-of-the-art open models,” https://blog.google/technology/developers/gemma-open-models/.

[23] "Gemma Model Card," https://www.kaggle.com/models/google/gemma.

All of the safety and ethical safeguards and interventions taken by AI companies such as those described above can relatively easily (and perhaps even secretly)[24] be removed from models with widely available weights. This is disturbing and means that companies' efforts to establish safeguards have to be understood as non-barriers for malicious actors.

Even without malicious intent to remove or subvert safeguards, it is still critical to perform continuous monitoring and de-biasing of foundation models with widely available model weights because fine-tuning and new use cases can impact the quality of outputs and the propensity for harmful bias. Moreover, unsecured models will not necessarily benefit from advances in the field and improvements in methodologies for mitigating the amplification of harmful bias.

We provided an assessment of developers' current practices in Appendix 4 of our AI Risk-Management Standards Profile for General-Purpose AI Systems and Foundation Models (https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf). We found no evidence of publicly communicated risk tolerances and varying capability assessments, with some developers focusing more on performance benchmarks than assessments focused on equity and representation.

It is critical that downstream developers are able to have adequate visibility into the data used, training processes, and output distribution to ensure their applications take these factors into consideration.

d. Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?

Unsecured models provide model developers with fewer mitigation measures to detect, interdict, and/or deter threat actors from taking malicious actions with these models. Advantages of AI systems typically include increased scale and/or increased speed. National security concerns we currently face by state and non-state actors can be increased in speed and scale due to widely available models. This could include increasing the volume of phishing messages or disinformation on social media, or automating steps in the cyber kill chain. Additionally, malicious actors who leverage AI might be able to create more sophisticated attacks than actors who do not use AI due to its ability to refine and assist with complicated tasks. We may not see big changes in the threat landscape or threat vectors (at least not soon), but we may see changes in adversaries' tactics, techniques, and procedures (TTPs). These changes in TTPs will be available to a broader set of actors due to widely available models. Traditional cost and knowledge barriers still exist (a modern laptop can barely run a 7B parameter model efficiently, requiring some computational resources to run even the smallest state-of-the-art unsecured models), and it is unknown at what point using unsecured models will provide an advantage

[24] “Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models,” https://arxiv.org/pdf/2310.02949.pdf.

over traditional methods, given how effective some malicious actions, such as cyberattacks, already are.[25] However, there are now examples of tailored unsecured models being used for hacking. Research evaluating the effectiveness of nine malicious LLM services available on the dark web found that the four most effective services were all based on open-source models (see Lin et al. 2024, “Malla: Demystifying Real-world Large Language Model Integrated Malicious Services”, https://arxiv.org/abs/2401.03315).

i. How do these risks compare to those associated with closed models?

Closed models, or more precisely, structured access such as through APIs, typically provide more opportunities than downloadable model weights allow, for mitigations such as content filters to prevent misuse, monitoring of usage to identify attempts at misuse (successful or unsuccessful), and system shutdown or rollbacks if problematic usage is identified. For example, OpenAI and Microsoft recently shared findings of state-affiliated threat actors using GPT-4 for assistance in cyber capabilities including writing content for phishing campaigns and exploring ways to avoid malware detection. Because the companies provide access through an API, they are able to monitor and study such misuses of their systems and take actions to prevent further harm, such as disabling accounts, terminating services, or limiting access to resources.[26]

ii. How do these risks compare to those associated with other types of software systems and information resources?

There are often many potential misuses of technologies, and there is a long history of identifying and mitigating risks of both open and closed source software. However, there can be an important difference between risks of misuse of software in sophisticated networked information systems, as compared with technologies in physical systems with less sophisticated networking, that sometimes seems overlooked: the feasibility and time requirements for fixing vulnerabilities and other risks after being newly identified. Often, vulnerabilities in software security are relatively easy and quick to patch; the norm is often public disclosure after 90 days. Vulnerabilities in embedded systems, biosecurity, and other cyber-physical systems are often harder and slower to patch; the norm is often secrecy or other control measures for years. Many of the vulnerabilities of large language models and other foundation models are imperfectly addressed via patches, as vulnerabilities such as susceptibility to data poisoning and prompt injection attacks are inherent to the models’ design and cannot be eliminated via patches.

The risks of foundation models are also sometimes compared to the risks of search engines (e.g., if you can Google how to make a bomb, does it matter if a model is willing to tell you?), but this comparison is limited for several reasons. Foundation models do not merely make pre-

[25] For example, ChatGPT was released in November 2022, and Zscaler reported a 47% rise in phishing attacks in 2022 alone. https://www.zscaler.com/blogs/security-research/2023-phishing-report-reveals-47-2-surge-phishing-attacks-last-year.

[26] “Disrupting malicious uses of AI by state-affiliated threat actors,” https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors.

existing information available to you, but can also identify new insights and pathways, and can be supplemented with tools and plugins to carry out actions in the world.[27] Importantly, because of how people relate to AI chatbots in particular, they can also be significantly more influential on people's decision making.[28] The comparison also neglects methods for amplifying the capabilities of base foundation models, such as the emerging class of foundation model-based agentic systems.[29]

e. What, if any, risks could result from differences in access to widely available models across different jurisdictions?

It seems prudent to assume that if a model is available in any jurisdiction, at least some malicious actors could and would distribute it across other jurisdictions. This also could reduce the effectiveness of sectoral-focused regulatory approaches, and increase the importance of horizontal or upstream requirements on foundation model developers and providers.

f. Which are the most severe, and which the most likely risks described in answering the questions above? How do these set of risks relate to each other, if at all?

The most severe would include: events impacting the health or human rights of large numbers of people, or disrupting critical societal systems or functions such as democratic institutions. The most likely, many of which are already occurring frequently, include: the spread of NCII including CSAM, and the promotion of harmful stereotypes, violence, and online radicalization. However, those categories are not mutually exclusive. Indeed, some not-yet-observed high-impact scenarios, such as large-scale disruptions to democratic institutions, could be seen as simply more-extreme versions of lower-impact scenarios that are already occurring, such as the spread of AI-generated synthetic media with the intention to deceive voters.

[27] "Building Bridges with Gorilla: Revolutionizing Language Models with APIs," https://opensourcecoder.medium.com/building-bridges-with-gorilla-revolutionizing-language-models-with-apis-ad81000c4fd1.

[28] See, e.g. "Exploring relationship development with social chatbots: A mixed-method study of replika," https://www.sciencedirect.com/science/article/pii/S0747563222004204.

[29] See, e.g., Cognition, "Introducing Devin, the first AI software engineer", https://www.cognition-labs.com/introducing-devin.

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

b. 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?

The state of the field of AI as we know it has greatly benefited from the open release of data, models, and their weights. This includes public release of models associated with pivotal research into bias in ML models, life-saving medical technologies, and more. The scrutiny made possible by the open release of model weights has resulted in critical bugs in these models being fixed[30], something that may not have occurred were these models closed.

As mentioned above under RFC Question 2a, one important issue with risks associated with widely available model weights is the framing of "safety". Models with widely available weights can provide benefits for increasing *safety for intended users*, in terms of reliability and security for those intended users. However, models with widely available weights also can pose substantial risks of reducing *safety for the general public* due to misuse by unintended users, in terms of potential for harmful misuse by malicious actors (e.g., for NCII, fraud, CBRN, or cyber attacks). These risks are likely to disproportionately harm women, minority groups, and vulnerable communities including children and the elderly.

We also note that the potential benefits to safety, security, and trustworthiness of AI from making model weights widely available sometimes rely on the free labor of open source communities.[31] In addition, as we mention under RFC Question 8a and elsewhere, benefits of broad independent evaluation for improving the safety, security, and trustworthiness of AI are not necessarily best supported by making model weights widely available. Those benefits can also be achieved by facilitating safe and protected independent researcher access.

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

Model evaluations are a critical tool to help determine the risks and benefits of foundation models in general, which can help inform the decision to make model weights widely available. Unfortunately, meta-analyses of model evaluations underscore significant limitations in the

[30] See, e.g., "Fixing All Gemma Bugs." https://unsloth.ai/blog/gemma-bugs.

[31] See, e.g., Widder et al. (2023), "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI," https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807.

current field of model evaluations including that the vast majority of evaluations focus on the text modality over others such as audio and video, and that the vast majority focus on capabilities instead of potential impacts on society.[32]

Evaluations of a foundation model's dangerous capabilities, including dual-use capabilities, are another critical part of assessing the risks of malicious misuse of a foundation model that have been relatively neglected; those malicious-misuse risks would tend to be compounded and harder to mitigate if and when the model weights of a dual-use foundation model are made widely available.

b. Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?

Currently, it is relatively simple and inexpensive to remove safeguards for models with widely available weights, e.g., via fine tuning, generation exploitation, and other "jailbreaking" techniques. (See, e.g., Gade et al. 2023, "BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B", https://arxiv.org/abs/2311.00117; Huang et al. 2023, "Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation", https://arxiv.org/abs/2310.06987; and Deng et al. 2023, "MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots", https://arxiv.org/abs/2307.08715).

In addition to investing in upstream protections to prevent a range of misuses of unsecured foundation models, we should invest in downstream protections to prevent specific misuses. However, we should not rely only on downstream protections such as requiring U.S. mail-order gene synthesis labs to screen orders as a way to prevent the creation of bioweapon agents. For example, for gene synthesis, it has long been recognized there are mail-order gene synthesis labs in China or elsewhere outside the United States that are much less likely than U.S. labs to follow standards for screening gene synthesis orders, which limits the value of mandatory screening of gene synthesis orders in the United States (see, e.g., Tucker 2010, "Double-Edged DNA: Preventing the Misuse of Gene Synthesis", https://issues.org/tucker-2/; Lovett 2024, "DNA bought online could be used to create dangerous pathogens, experts warn", https://www.telegraph.co.uk/global-health/science-and-disease/dna-could-be-bought-online-to-create-dangerous-pathogens/).

[32] "Sociotechnical Safety Evaluation of Generative AI Systems," https://arxiv.org/abs/2310.11986 and "Evaluating the Social Impact of Generative AI Systems in Systems and Society," https://arxiv.org/abs/2306.05949.

d. Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?

No, there are no effective ways to achieve this that we know of. Foundation model developers that release a model's parameter weights via open-source, fully open access, or downloadable access, and developers that suffer a leak of model weights, will be unable to effectively regain control, restrict access, or decommission those models or derivative AI systems that others build using those released or leaked foundation model weights. Once model weights become widely available, it also becomes impossible for critical security updates to be effectively propagated to all instances of that model.

e. What if any secure storage techniques or practices could be considered necessary to prevent unintentional distribution of model weights?

If a model developer aims to use information security to prevent the release of model weights, standard information security controls can and should be used. As recommended in Measure 2.7 of our AI Risk-Management Standards Profile for General-Purpose AI Systems and Foundation Models (https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf), in order to protect the integrity and confidentiality of model parameter weights, model developers should implement the NIST Cybersecurity Framework, or an approximate equivalent such as NIST SP 800-171 or ISO/IEC 27001, with at least the following security controls or their approximate equivalents:

- For frontier models: High-value asset guidance (e.g., per NIST SP 800-171 and NIST SP 800-172), or high-impact system baseline per NIST SP 800-53B as an informative reference for the NIST Cybersecurity Framework, or approximate equivalent
- For foundation models: Moderate-impact system baseline guidance (e.g., per NIST SP 800-171), or moderate-impact system baseline per NIST SP 800-53B as an informative reference for the NIST Cybersecurity Framework, or approximate equivalent

However, currently available information security practices may not be sufficiently robust to prevent access and exfiltration of model weights by APTs and other state-level adversaries (see, e.g., the October 2023 RAND interim report "Securing Artificial Model Weights", https://www.rand.org/pubs/working_papers/WRA2849-1.html)

7. 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?

a. What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?

The U.S. government should restrict the availability of dual-use foundation model weights in cases where particular dual-use models would be substantially more useful for CBRN, cyber attacks, or other attacks with the potential for large-scale harm to the public. No open models have been shown to significantly add risk in these domains at this point in time. However, as model capabilities increase, this could potentially become a concern, at which point the government will want to consider regulatory action.

One avenue could be the export controls in the Export Administration Regulations (EAR). The Department of Commerce could amend parts of the EAR, such as a category of the Commerce Control List (CCL), to clearly indicate that dual-use foundation models would be subject to the EAR. The Department of Commerce could also amend §734.7 to state that making a dual use foundation model's weights widely available should not be regarded as a "publication" exception to the EAR. (§734.7(b) and §734.7(c) give precedents for listing some types of technologies, such as software instructions for producing a firearm using additive manufacturing equipment, for which "publication" does not provide an exception to the EAR.) For discussion of related issues, see, e.g., the working paper Bloomfield 2024, "Export Controls and Artificial Intelligence Biosecurity Risks", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4741033.

At some point, technical advances may enable cryptographic software- and hardware-based "proof of safety" checks, such that particular models would not run without assurance of satisfying particular safety criteria. However, feasibility and availability of such options is not assured, and would first require a large amount of research and investment. (See, e.g., Russell 2023, "Stuart Russell Testifies on AI Regulation at U.S. Senate Hearing", https://humancompatible.ai/blog/2023/09/11/stuart-russell-testifies-on-ai-regulation-at-u-s-senate-hearing/.)

b. How might the wide availability of open foundation model weights facilitate, or else frustrate, government action in AI regulation?

From a regulator's perspective, unsecured foundation models currently represent the biggest challenge to government action because of their horizontal nature (foundation model) and their wide accessibility (open), making it difficult to define both the right scope for the governance framework and the appropriate target population for government action. It is likely outside of the capability of any public agency charged with supervising the implementation of rules for AI models to keep an eye on anyone who can now download and modify the model. If model weights are widely available, this raises the question of at which point a “user” who adapts the model to their own use case becomes a developer who has to disclose how they changed the model and which new risks might arise with the changes. The debates surrounding requirements for open source AI models in the European Union’s AI Act illustrate these challenges.

Under the EU’s AI Act, open source GPAI models or foundation models are exempted from some foundation-model regulatory requirements, such as providing a public summary of training data. However, open GPAI models need to fully comply with regulatory requirements if they are considered to pose systemic risks. One criterion for determining a model poses a systemic risk is “high impact capabilities”, as indicated using metrics or proxies such as “if the cumulative amount of compute used for its training, measured in floating point operations (FLOPs), is greater than 10^25” (Article 51, AI Act, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf). Additionally, the European Commission reserves the right to adapt this threshold following technological developments and to classify models as having systemic risk if they meet additional criteria or benchmarks (Article 51). Combining the threshold definition with additional benchmarks grants the European regulator flexibility to adapt to a changing technological landscape.

However, this approach also leaves some uncertainty for developers and users of open source models regarding whether or not their models will benefit from the exemptions because the Commission can classify them as high-risk on (currently) undefined metrics and benchmarks. In the absence of further clarification, developers seem likely to default to the implementation of the most stringent rules instead of claiming exemptions, to ensure compliance (see, e.g., Burri 2023, “A challenge for the law and artificial intelligence”, https://www.nature.com/articles/s42256-023-00768-5). Despite these drawbacks, superior approaches to defining systemic risk do not yet appear to be available. Moreover, this approach should be viable provided the enforcing agency is reasonably well equipped with expertise and resources and clarifies the additional indicators and benchmarks in due time.

The European regulators’ approach does not address the specific challenge generated by open foundation models which is that developers relinquish control over model use once released (Kapoor et al. 2024, “On the Societal Impact of Open Foundation Models”, https://crfm.stanford.edu/open-fms/paper.pdf). Which rules apply to models that have been modified after release remains an open question, e.g., if additional or renewed requirements for the evaluation of risks are needed and who would need to provide documentation.

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 establishing common standards, drawing upon the best available research and practices. Many aspects of such standards should be risk-based, and risk management standards should apply to both unsecured and secured models without preventing the eventual open release of a model if its risks are sufficiently low. As we discuss below in our response to NTIA RFC question 8a, we recommend that foundation model developers that plan to provide open access to their model weights should use a staged-release approach with initial structured-access release to monitor for novel risks and harms without releasing model weights, 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.

However, currently, there is limited accountability when developers fail to follow best practices or when harmful impacts materialize. Thus, the U.S. government should do more than just set soft-law standards: there is also a need for hard-law regulatory requirements with enforcement by appropriate regulators. Government agencies and departments should be provided with sufficient resources to uphold pre-existing laws in the age of AI proliferation. Where gaps are identified, additional regulatory authority or new federal AI laws with provisions for sufficient oversight and fines or other penalties for irresponsible actions would also be highly valuable. Without such federal laws and enforcement, companies may perceive net incentives to move too hastily to develop and deploy excessively risky AI systems.

For more recommendations and related discussion, see our policy brief, https://cltc.berkeley.edu/publication/policy-brief-on-ai-risk-management-standards-for-general-purpose-ai-systems-gpais-and-foundation-models/.

i. Should other government or non-government bodies, currently existing or not, support the government in this role? Should this vary by sector?

If foundation models released in one sector could be easily taken and re-trained for different sectors (as mentioned earlier in this document), a horizontal approach seems more appropriate for foundation models than only a sector-specific approach.

g. What should the U.S. prioritize in working with other countries on this topic, and which countries are most important to work with?

The United States should aim for harmonization, or at least interoperability, of regulatory requirements related to risk management for models with widely available weights. One of the most important areas is the EU, because of the EU AI Act's hard-law regulatory approach to foundation models (and specifically open foundation models), as well as the existence of the U.S.-EU Trade and Technology Council (TTC). The United States should also continue to engage with the Organisation for Economic Co-operation and Development (OECD) and the United Nations among other intergovernmental fora to support global standards and frameworks.

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?

Available evidence has demonstrated harmful impacts from misuse of unsecured models, in particular for child sexual abuse material (CSAM), non-consensual intimate imagery (NCII), disinformation at scale, facilitating cyberattacks, enabling online radicalization, and promoting harmful stereotypes and violence. Some have argued that the marginal risk of open foundation models is relatively minimal in at least some cases.[33] The notion of “marginal risk” can be a helpful framing to ground the discussion of risks in a broader context. However, some of the most well-documented risks, such as the promotion of harmful stereotypes and violence, were not included in that particular study.

In addition, risk management should include some margin of safety in the face of uncertainty about the capabilities of a model, or of AI systems incorporating a model. There is potential for surprise in dangerous capabilities: A model’s capabilities could be substantially greater than expected at the time of the model’s release.[34]

One important risk of frontier models is that some aspects of their capabilities often are not identified until months or years after a model’s release. Risk management also should consider factors other than the base model, such as scaffolding, plugins, agentic wrappers, etc., as these could substantially increase the effective capabilities of AI systems incorporating models. These risks could be further compounded by increasingly sophisticated robotics systems, such as for biological laboratories, that could reduce the digital-to-physical divide that is currently among important barriers to physical harm from AI systems.

a. How should these potentially competing interests of innovation, competition, and security be addressed or balanced?

There is broad agreement, including among advocates of open source foundation models, on key points including that some models should not be open source; and the most broadly capable foundation models (including dual-use foundation models or other “frontier” models) pose a higher risk.

There are disagreements on other key points, including the degree of risk posed by frontier models (e.g. on whether LLMs could provide information that would be more useful to bioterrorists than information on the internet); and the default assumptions for methods of release (e.g. fully open vs. fully closed or something in between).

[33] See, e.g. the marginal-risk framework of Kapoor et al. (2024), “On the Societal Impact of Open Foundation Models”, https://crfm.stanford.edu/open-fms/paper.pdf.

[34] Kapoor et al. (2024) note that their marginal-risk framework does not account for unknown risks of open model releases.

In addition, many of the benefits of open source, 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.

We recommend middle-ground alternatives, rather than accepting a false dichotomy of "always fully open" vs. "always fully closed", highlighting the following two options:

1. Staged release, including a phase with structured access such as through server-hosted APIs. Foundation model developers that publicly release the model parameter weights for their foundation 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. This is a consideration that should be weighed against the benefits of unsecured models, 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.

   Therefore, we recommend the following, drawing upon our Profile and policy brief: 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 secured 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.[35] The largest-scale or most capable models (including dual-use foundation models as defined in Executive Order 14110) 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.

   To make this recommendation less burdensome to implement, we suggest resource-constrained foundation model developers such as academics make use of third-party model hosting infrastructure services (e.g. Amazon SageMaker), especially for models far below dual-use foundation model thresholds. We suggest that infrastructure service providers, in turn, continue to develop their services to take more labor-intensive aspects

[35] We are not aware of a specific case where a foundation model developer has integrated all aspects of this recommendation. However, there are numerous examples of parts of our recommendation. For example, OpenAI's GPT-2 is an example of a staged release with several partial, increasingly capable, downloadable-weights versions of the model over the course of 2019, culminating in the full GPT-2 version release in November 2019 (see, e.g., "Better language models and their implications", https://openai.com/research/better-language-models; "GPT-2: 1.5B release", https://openai.com/research/gpt-2-1-5b-release). By contrast, GPT-3 and GPT-4 are available through a server-hosted API with monitoring for indications of misuse, but those models have not become available with downloadable weights (see, e.g., "Lessons learned on language model safety and misuse", https://openai.com/research/language-model-safety-and-misuse).

of secure model hosting, e.g. abuse monitoring, off of the shoulders of model developers.

2. More freely-accessible secured foundation models. When models are available cost-free (or cost-subsidized), but still provided via structured access such as behind APIs and hosted UIs, they could achieve many of the potential benefits of widely available model weights that NTIA noted in this RFC, such as fostering growth among less resourced actors, and helping to widely share access to AI's benefits. Moreover, they could achieve those benefits without giving up the risk-mitigation benefits of structured access, such as monitoring server traffic and logs for indicators of abuse and other important risks.

   Some freely-accessible secured models such as OpenAI's ChatGPT 3.5 already exist, but more advanced models such as GPT-4 tend to be only available at prices that can exclude less-resourced actors. We recommend taking steps to promote the cost-free (or cost-subsidized) availability of more advanced foundation models to everyone, or at least to key less-resourced actors such as small businesses, academic institutions, underfunded entrepreneurs, nonprofits, low-income individuals, etc.

   In addition to removing monetary barriers, freely-accessible secured models can often be more inclusive to less technologically sophisticated users. This is because, for secured models, the model provider can provide a hosted user interface. On the other hand, for models where weights are distributed, a degree of technical expertise is required to download model weights, host the model, and learn how to query the model programmatically via an API or to find open-source UI options that are compatible with the model in question.

   There are various ways that the government could improve the accessibility of secured foundation models. A couple of preliminary suggestions follow, though the list is not comprehensive, and further analysis is needed:

   (a) One approach would be requiring foundation model developers to provide free or cost-subsidized model access to less-resourced actors. This sort of mandate could be funded or unfunded by the government. The benefits of this approach include that it leverages existing infrastructure and expertise of free-market foundation model developers. A drawback of this approach is that it continues to drive data centralization and other forms of power concentration in the hands of a few technology companies.

   (b) Another approach the government could take to improve freely-accessible models would be to develop or commission the development of public versions of models. This could look like having a government lab or government-contracted lab that trains its own models designed to emulate popular paid AI models on the private market. For example, the government would attempt to develop its own ChatGPT-like system and provide it for free to low-income individuals, nonprofits, small businesses, etc. The benefits and drawbacks of this approach are the opposite of approach (a). That is, it could help reduce the centralization of data and power in the hands of a few technology companies, but it would be non-trivial for the government to match the quality of leading private models and to learn how to monitor their usage effectively.

Similar discussion and recommendations can also be found in others' policy analyses such as "Open-Sourcing Highly Capable Foundation Models" (https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models) and industry guidance including the Partnership on AI's "Guidance for Safe Foundation Model Deployment" (https://partnershiponai.org/modeldeployment/).

Moreover, the above-mentioned approaches could be interoperable with the tiered regulatory obligations for providers of open foundation models under the EU AI Act. As we mentioned in our response to question 7b, the regulatory requirements for open foundation models are greater for models posing systemic risks, e.g., if the model training uses greater than 10^25 FLOPs of compute.

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?

One of the challenges is that the quantity (floating point operations or FLOPs) of compute used in model training is one of the few, relatively objective pre-deployment indicators of potential model capabilities. Model evaluations are also useful but less accurate and objective, at least for now. Post-deployment impact metrics also are important but are not available pre-deployment. Thus, model training compute is an important pre-deployment risk indicator, and should continue to be used as such, at least for the time being. However, other relevant metrics such as model evaluations also should be incorporated as they become more reliable and practical. The European Commission plans to take a similar approach to identifying models posing systemic risks, as we mentioned previously. (For related discussion, see, e.g., Bommasani 2023, "Drawing Lines: Tiers for Foundation Models", https://crfm.stanford.edu/2023/11/18/tiers.html; we also list factors to consider as part of foundation model impact assessment or risk assessment in our 2023 AI Risk-Management Standards Profile for General-Purpose AI Systems and Foundation Models, https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf, largely under either Map 1.1, Map 5.1, or Measure 2 in the Profile.)