1. How should the NTIA define "open" or "widely available" when thinking about foundation models and model weights?
    a. Is there evidence or historical examples suggesting that weights of models similar to currently-closed AI systems will, or will not, likely become widely available? If so, what are they?

        Yes. Typically, there is a delay between the time at which a new AI system is demonstrated ("demoed") and the time at which the weights of a similar AI system are made available. As an example, on OpenAI released a paper documenting their GPT-3 language model on May 28, 2020 and opened up a (closed) beta for users to interact with it on June 11, 2020. It took roughly 2 years for weights of models of similar performance (like Meta's OPT models: https://arxiv.org/abs/2205.01068) to be made public, and even then it was done in a way intended to limit access. Similarly, high-quality neural speech synthesis software like WaveNet (https://deepmind.google/technologies/wavenet/) and Tacotron (https://arxiv.org/abs/1703.10135) were first demoed nearly a decade ago and closed voice cloning models were developed by academics several years ago (https://arxiv.org/abs/1711.11293), but only in the past year have we seen widespread availability & use of open-weight models for high-quality voice cloning.

        There are very few examples of weights from closed AI systems being leaked and becoming unintentionally widely available. Even some of the releases reported as "leaks" (like Meta's original LLaMA models) were knowingly released to a wide set of researchers without much restriction so that they could be widely used. The most salient example of a true leak is Mistral's "Miqu" model leak (https://venturebeat.com/ai/mistral-ceo-confirms-leak-of-new-open-source-ai-model-nearing-gpt-4-performance/), which when released was not comparable in quality to alternative models (both open and closed).

    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?

        Not generally. I would estimate a normal lag between the first closed model being demonstrated at a given capability level and the first open model with similar capabilities to be between 6 months - 4 years.

For types of foundation models that have widespread appeal—like chatbots—I would expect the timeframe to be relatively short (perhaps less than 2 years), whereas for those with narrow appeal—like nuclear physics—I would expect the timeframe to be relatively long (measured in several years). It is possible to use estimates of "algorithmic progress" (like https://epochai.org/blog/algorithmic-progress-in-language-models)–or the speed at which the computational cost to reach a particular bar of performance decreases–to estimate this. But such estimates do not factor in the growth in *spending* that may occur. It seems likely to me that, as the scale of compute required to reach impressive performance goes up, there will be fewer and fewer actors who can afford to participate in development.

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?"

No comment.

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?

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

In my view, one promising prospective form of access is a "source-available weights + closed deployment" scheme as Cohere has done with their Command-R model (https://txt.cohere.com/command-r/). Cohere released the weights of the model so that researchers & hobbyists can inspect and evaluate them, but with licensing that restricts other entities from legally deploying them. This allows Cohere to manage how and by whom their models are deployed, whether within their own infrastructure or via licensing agreements with other trusted organizations. If any problems with the model are discovered as the model is used in the real-world, Cohere as a single actor can mitigate them by releasing a new version, without stifling research on the original model weights. "Source-available" is a known strategy, with other major projects such as Epic Games' Unreal Engine opting for a similar approach.

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?

My only comment here is that nearly all of the risks associated with widely available model weights are downstream of the fact that since these weights are software/data, they can be copied easily and thus cannot be systematically recalled in the way that, say, a defective automobile can be recalled.

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

Other commenters will likely write better on this. No comment.

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

It is unlikely that all actors training foundation models will be diligent in removing personally-identifiable information (PII) from the training datasets used to create their models. This seems especially unlikely as the compute requirements to produce usable models decreases over time. Therefore, it seems likely that there will exist widely-available model weight that contain significant amounts of PII. If one was actively trying, one could extract this PII from those models given partial cues. For example, if the training dataset contained leaked Social Security Numbers (SSNs), it is possible that when the model is prompted with a few known name + SSN pairs plus a real leaked name, it may respond with the corresponding leaked SSN. However, this will be true of both widely available model weights and closed-weight models. The only way to mitigate this is would look like requiring some form of redaction of PII in training datasets.

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?

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

Other commenters will likely write better on this. No comment.

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

Other commenters will likely write better on this. No comment.

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

Other commenters will likely write better on this. No comment.

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?

Other commenters will likely write better on this. No comment.

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

Other commenters will likely write better on this. No comment.

    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?

Widely-available model weights have been a primary enabler for research on privacy & security, model internals (or “interpretability”), learning dynamics, ethics, and alignment. Here is a table containing a non-exhaustative set of examples:

Date	Impact Category	Impact Summary	Requires Weights?	Requires Data?	Base Model(s)	Source Link
September 9, 2023	Privacy and Security		Yes	No	GPT-2, Bloom, Pythia	Reverse-Engineering Decoding Strategies Given Blackbox Access to a Language Generation System
November 18, 2021	Privacy and Security	Enabled systematic evaluation of linguistic novelty in text generation, quantifying the extent to which language models copy from their training data. This provides insights into privacy risks arising from memorization.	Yes	Yes	GPT-2	How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN
January 19, 2022	Privacy and Security	Examines how GPT-J can solve CS 101 assignments and verifies that its not copying from its training data	Yes	Yes	GPT-J	Fooling MOSS Detection with Pretrained Language Models
February 10, 2022	Model Internals	Enabled development of new method to locate & edit factual associations in neural networks.	Yes	No	GPT-2, GPT-J	Locating and Editing Factual Associations in GPT
February 15, 2022	Privacy and Security	A large scale and systematic analysis of memorization in large language models.	Yes	Yes	GPT-Neo, GPT-J, GPT-NeoX, T5	Quantifying Memorization Across Neural Language Models
October 13, 2022	Model Internals	Enabled development of more scalable fact-editing methods for neural networks, to adjust 1000s of associations at once.	Yes	No	GPT-J, GPT-NeoX	Mass-Editing Memory in a Transformer

March 14, 2023	Model Internals	Enabled more precise investigation and hypothesis-testing on how LLMs form predictions across their layers.	Yes	Yes	Pythia, GPT-Neo, BLOOM, GPT-2, OPT, LLaMA	Eliciting Latent Predictions from Transformers with the Tuned Lens
March 20, 2023	Ethics	Enabled an analysis of gender, ethnicity, and profession biases in text-to-image models by using CLIP embeddings as a common representation space. Developed novel techniques leveraging CLIP to assist labeling.	Yes	No	Stable Diffusion, CLIP	Stable Bias: Analyzing Societal Representations in Diffusion Models
May 11, 2023	Ethics	Investigated toxicity and quality filters for pretraining, the effects of pretraining scrape time, and composition of pretraining data on many benchmarks	Checkpoints	Yes	Custom trained models	A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity
May 24, 2023	Privacy and Security		Yes	No	GPT-2, LLaMA, Vicuna	Adversarial Demonstration Attacks on Large Language Models
May 31, 2023	Privacy and Security	Studies whether which data will be memorized by a LLM can be predicted before the LLM is fully trained	Checkpoints	Yes	Pythia	Emergent and Predictable Memorization in Large Language Models
June 6, 2023	Model Internals	Enabled validation of a method that provably erases certain kinds of information from neural networks, allowing for "surgical" removal of undesirable	Yes	No	Pythia, LLaMA	LEACE: Perfect linear concept erasure in closed form

		behaviors.				
June 7, 2023	Privacy and Security	Enabled generation of watermarked text to test the robustness of various "paraphrasing attacks".	Yes	No	LLaMA	On the Reliability of Watermarks for Large Language Models
June 30, 2023	Alignment		Yes	No	Many	Stay on topic with Classifier-Free Guidance
June 30, 2023	Privacy and Security		Checkpoints	Yes	Custom trained models	Measuring Forgetting of Memorized Training Examples
July 13, 2023	Learning Dynamics		Yes	Yes	Pythia	Layer-wise Linear Mode Connectivity
July 27, 2023	Privacy and Security	Enabled discovery, testing, and responsible disclosure of adversarial suffixes that universally "jailbreak" LLMs, including proprietary models. Would have been impossible to discover without access to open-weight models.	Yes	No	LLaMA, Llama2, MPT, Pythia, Falcon	Universal and Transferable Adversarial Attacks on Aligned Language Models
August 1, 2023	Ethics		Yes	Yes	Stable Diffusion	The Bias Amplification Paradox in Text-to-Image Generation
August 2, 2023	Privacy and Security		Checkpoints	Yes	Pythia	Tools for Verifying Neural Models’ Training Data
August 8, 2023	Ethics		Yes	Yes	Pythia	SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
August 20, 2023	Model Internals	Enabled application of lightweight editing method for controlling model computations, and investigation of scaling behavior.	Yes	No	GPT-2, OPT	Activation Addition: Steering Language Models Without Optimization

August 24, 2023	Other	Surveys illicit uses of open and closed models	Yes	No	Many	Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities
August 25, 2023	Model Internals	Enabled development of improved method for editing and erasing targeted concepts from text-image models.	Yes	No	Stable Diffusion	Unified Concept Editing in Diffusion Models
August 29, 2023	Learning Dynamics		Checkpoints	Yes	Custom trained models, Pythia	Characterizing Learning Curves During Language Model Pre-Training: Learning, Forgetting, and Stability
August 30, 2023	Privacy and Security	Enabled development and testing of much stronger red-teaming method against both open and proprietary LLMs.	Yes	No	LLaMA, Llama2	Red-Teaming Large Language Models using Chain of Utterances for Safety-Alignment
September 9, 2023	Privacy and Security	Reverse engineers the "top-p" and "top-k" parameters used when generating text from a LLM	Yes	No	GPT-2, BLOOM, Pythia	Reverse-Engineering Decoding Strategies Given Blackbox Access to a Language Generation System
September 15, 2023	Model Internals	Enabled discovery of an interpretable basis for cataloging features learned by LLMs.	Yes	Yes	Pythia	Sparse Autoencoders Find Highly Interpretable Features in Language Models
September 23, 2023	Privacy and Security		Yes	Yes	Many	From Text to Source: Results in Detecting Large Language Model-Generated Content
September 26, 2023	Privacy and Security	Enabled testing the robustness of a proposed method for automated lie detection across different LLMs.	Yes	No	LLaMA	How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions
September 28, 2023	Alignment		Yes	No	Pythia	Beyond Reverse KL: Generalizing Direct Preference Optimization

						with Diverse Divergence Constraints
September 29, 2023	Privacy and Security	Enabled discovery that previous model editing methods leave sensitive information that an attacker can then extract.	Yes	No	GPT-J, Llama2, GPT-2	Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks
September 29, 2023	Alignment		Yes	No	CodeGen, Pythia, MPT	Benchmarks for Detecting Measurement Tampering
October 1, 2023	Learning Dynamics		Checkpoints	No	OPT, Pythia	JoMA: Demystifying Multilayer Transformers via JOint Dynamics of MLP and Attention
October 3, 2023	Model Internals	Enabled efficient investigation and analysis of inner workings and activation patterns of neurons in transformer MLP layers through easy access to state-of-the-art interpretability data.	Yes	No	GPT-2, Pythia	DeepDecipher: Accessing and Investigating Neuron Activation in Large Language Models
October 4, 2023	Alignment	Enabled exploration of promising methods to combat "overoptimization" problem when performing safety tuning on models. Both the tuned and supervisor models were open-weight.	Yes	Yes	Pythia, LLaMA	Reward Model Ensembles Help Mitigate Overoptimization
October 5, 2023	Alignment	Enabled discovery and investigation of a flaw in standard safety tuning methods that incentives long responses instead of just incentivizing	Yes	No	LLaMA	A Long Way to Go: Investigating Length Correlations in RLHF

		helpfulness.				
October 10, 2023	Other		Yes	No	LLaMA-2	Representation Engineering: A Top-Down Approach to AI Transparency
October 10, 2023	Privacy and Security	Enabled proof-of-concept demonstrating the recovery of sensitive information from content embeddings, with implications to the secure design of multi-user semantic search applications.	Yes	No	T5	Text Embeddings Reveal (Almost) As Much As Text
October 11, 2023	Ethics	Presents a cheap and efficient technique to intervene on LMs to decrease toxicity in generations	Yes	No	GPT-2, OPT, Pythia	Goodtriever: Adaptive Toxicity Mitigation with Retrieval-augmented Models
October 12, 2023	Alignment		Yes	No	Pythia	Interpreting Reward Models in RLHF-Tuned Language Models Using Sparse Autoencoders
October 17, 2023	Alignment	Enabled investigation of more specific & scalable safety tuning method, incorporating multidimensional preferences.	Yes	No	LLaMA	Preference Ranking Optimization for Human Alignment
October 17, 2023	Privacy and Security	Enabled testing of quantization-based watermarking methods.	Yes	No	LLaMA, GPT-Neo	Watermarking LLMs with Weight Quantization
October 19, 2023	Ethics		Yes	No	GPT-2	Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model
October 19, 2023	Alignment	Enabled testing of improved RLHF algorithm that uses explicit constraints.	Yes	No	LLaMA	Safe RLHF: Safe Reinforcement Learning from Human Feedback

October 19, 2023	Alignment	Enabled showing that existing vision-language models can be reused as supervisors for reinforcement learning, and that larger models are more robust at this.	Yes	No	CLIP	Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning
October 22, 2023	Privacy and Security		Checkpoints	Yes	Pythia	MoPe: Model Perturbation-based Privacy Attacks on Language Models
October 23, 2023	Model Internals	Enabled investigation of how input-output functions are contextually triggered in LLMs, and how they can then be triggered in a targeted way post-hoc.	Yes	No	GPT-J, GPT-NeoX, Llama2	Function Vectors in Large Language Models
October 23, 2023	Learning Dynamics		Checkpoints	Yes	Pythia	Meta- (out-of-context) learning in neural networks
October 23, 2023	Model Internals		Yes	Yes	Pythia	Understanding the Inner Workings of Language Models Through Representation Dissimilarity
October 24, 2023	Model Internals		Yes	Yes	GPT-2, Pythia	Characterizing Mechanisms for Factual Recall in Language Models
October 24, 2023	Model Internals		Yes	No	GPT-J, Pythia, LLaMA	In-Context Learning Creates Task Vectors
October 25, 2023	Learning Dynamics		Checkpoints	No	MultiBERTs	Subspace Chronicles: How Linguistic Information Emerges, Shifts and Interacts during Language Model Training
October 26, 2023	Model Internals		Yes	No	Pythia	Codebook Features: Sparse and Discrete Interpretability for Neural

						Networks
October 31, 2023	Model Internals		Yes	No	BERT, GPT-2, OPT	Large Language Models Converge on Brain-Like Word Representations
November 28, 2023	Privacy and Security		Yes	Yes	GPT-Neo, Pythia, RedPajama-INCITE	Scalable Extraction of Training Data from (Production) Language Models
March 19, 2024	Other	Measures the discrepancy between reported and actual model knowledge cutoffs (primarily through investigating training data)	Yes	Yes	Pythia, OLMo, RedPajama	Dated Data: Tracing Knowledge Cutoffs in Large Language Models

c. 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.)?

Other commenters will likely write better on this. No comment.

d. How can the diffusion of AI models with widely available weights support the United States' national security interests? How could it interfere with, or further the enjoyment and protection of human rights within and outside of the United States?

Other commenters will likely write better on this. No comment.

e. How do these benefits change, if at all, when the training data or the associated source code of the model is simultaneously widely available?

Wide availability of training data or associated source code is helpful but not required for the benefits of foundation models with widely available weights to materialize. In the table above you can see that much beneficial research, modification, and use is possible without access to original training data. However, when training data and source code are available, it becomes more feasible to predict the behavior of the model, to modify it in targeted ways, and to study learning dynamics of 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.

- **Dataset preprocessing pipelines - Even in the event that the original training dataset cannot or should not be released, it may be beneficial for others to be able to re-use the logic that was used to curate the training datasets. By doing so, institutional actors can coordinate on best practices for dataset cleaning / management and researchers can have some indication of what kind of data a model was trained on.**
- **Runtime safety filters - Systems like Meta's Llama Guard (https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/) and OpenAI's free-to-use moderation API (https://platform.openai.com/docs/guides/moderation) are designed to help developers guard their systems against misuse/abuse. Because they are separate from the underlying foundation model, anyone can use them no matter which model they are building on. If widely available, different actors can coordinate to implement the same best-practices for security, privacy, and anti-fraud.**
- **"Reward models" - In order to convert a foundation model into a chatbot, model developers will sometimes train an auxiliary model (called a "reward model") to score outputs based on qualitative attributes like toxicity & response length. These are then used to further tune the foundation model to behave in a way the model deployer considers appropriate, rather than behaving as it would if trained only on the original corpus. Access to these reward models prior to the release of a model (either open-source or closed-source) may allow independent researchers, auditors, and other teams to predict extreme or unwanted behaviors before deployment better than they would be able to if they only had access to the original "base" foundation model. Organizations were originally slow to release these models, but they are useful for downstream applications and we are beginning to see developers release / evaluate them publicly. See AllenAI's Reward Bench: https://github.com/allenai/reward-bench**

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?

I do not think that model evaluations can help much in determining the *general* risks or benefits associated with making weights of a foundation model widely available. For *specific* capabilities of concern, it may be possible to use model evaluations to test how easy it is to elicit those capabilities from a particular model. However, this is really only possible when given *full* access to the model, including the ability to finetune its weights and combine it with other systems. If this access is not granted, there is a potential that model evaluations will underestimate the risks or benefits from a particular system.

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?

I believe there are already established practices around securing corporate, medical, and military information that would be applicable to this, and that would do well at ensuring that model weights are only made available when the developer intends to.

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

The prospects appear strong to me.

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?

It may be possible to limit use of weights from a particular open foundation model through policies to restrict cloud model inference services from hosting that model and any known derivatives from it. This would not be very costly to implement, as it is essentially a "model blacklist", but it would push actors who intended to use that model towards local/self-hosted inference or hosted inference outside the USA. At present, model inference is likely very concentrated, with most users of foundation models relying on a hosting service rather than managing their own infrastructure.

There is no way to regain control over the weights of an open foundation model that, either inadvertently or purposely, have already become widely available. In fact, it is likely that doing so will encourage the further distribution of those weights, due to the Streisant Effect (https://en.wikipedia.org/wiki/Streisand_effect).

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

This is not my area of expertise.

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.

As stated earlier, a comprehensive evaluation of the model would call for inspection by the evaluator/certifier/red-team of the entire model, including weights, training data, and other code. It may also call for disabling safety checks and allowing them to further train the model a bit, to understand what the consequences would be if a malicious user were to somehow disable runtime checks and access its underlying abilities. However, if red teaming becomes a pre-condition for deployment of models, developers are incentivized to keep the information exchanged with evaluators to a minimum, only giving them access to a locked-down API to make queries on.

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?

It is definitely possible to do empirical tests to see how the model defined by a set of weights behaves across a range of contexts. One example testbed—which I had a very minor hand in developing—that is used for large language models (LLMs) is EleutherAI's "lm-evaluation-harness" (https://github.com/EleutherAI/lm-evaluation-harness). There are also evaluation organizations like METR (https://metr.org/) and Apollo Research (https://www.apolloresearch.ai/) that are working on systematic testing of AI models and infrastructure for it.

With respect to verification, it is possible to use tools from cryptography to verify that a set of model weights has certain properties. For example, it is possible to check that the file containing a set of model weights corresponds to some public checksum, so a user can know they are downloading the correct model. It is also possible to do more advanced forms of verification & testing. For example, it is possible to modify model weights to plant undetectable backdoors in them (https://arxiv.org/abs/2204.06974), and such backdoors are suspected to be very hard to remove without knowledge of the watermarking entity's keys (https://arxiv.org/abs/2401.05566). These could likely be used to track the provenance of weights from foundation models, and to attach markers that indicate that they have gone through a verification process. I would recommend consulting

with cybersecurity & cryptography experts on these matters. I suspect, however, that even assuming they are feasible to implement, these protocols are not standardized and would need to gain buy-in in order to implement.

6. What are the legal or business issues or effects related to open foundation models?
    a. In which ways is open-source software policy analogous (or not) to the availability of model weights? Are there lessons we can learn from the history and ecosystem of open-source software, open data, and other "open" initiatives for open foundation models, particularly the availability of model weights?

        Model weights are software, just as with other software. However, the way that model weights are created does not primarily involve "writing source code" in the same way as with ordinary software. Also, the process of going from the original behavioral specification (the training data) to running code (the trained model) is significantly longer and more expensive. This changes some but not all of the dynamics. The Open Source Initiative has been discussing this matter and should have relevant institutional knowledge. Assuming that they are also submitting comments, I would defer to them on possible lessons from history.

    b. How, if at all, does the wide availability of model weights change the competition dynamics in the broader economy, specifically looking at industries such as but not limited to healthcare, marketing, and education?

        Wide availability of model weights may enable sectors (such as healthcare and education, in some cases) that are otherwise prohibited from sending data externally to benefit from AI developments.

    c. How, if at all, do intellectual property-related issues—such as the license terms under which foundation model weights are made publicly available—influence competition, benefits, and risks? Which licenses are most prominent in the context of making model weights widely available? What are the tradeoffs associated with each of these licenses?

        There have been many many different licenses applied to foundation model weights. For actors intending to use the model in a commercial fashion, the most common and most preferable licenses are MIT and Apache, as they are quite permissive in what they allow. The tradeoff is that there is no ability to restrict how the model is used under these licenses. Several organizations like Meta and NVIDIA have opted for more restrictive custom licenses that restrict the ways their models are used, or the scale of use. Some of these are intended to prevent misuse and illegal use by bad actors. Others are intended to reduce competition or to give an advantage to the company that developed

the model. For example, the license under which Meta released their Llama2 models prevents the model from being used by enterprises larger than a particular size, which limits who can compete in serving that model. (https://www.theverge.com/2023/10/30/23935587/meta-generative-ai-models-open-source) For academic research, the license does not *necessarily* make a difference, but it changes the set of organizations that are willing to host the model and the number of users that are ultimately exposed to it, which can have downstream impacts on which models researchers are interested in investigating.

d. Are there concerns about potential barriers to interoperability stemming from different incompatible "open" licenses, e.g., licenses with conflicting requirements, applied to AI components? Would standardizing license terms specifically for foundation model weights be beneficial? Are there particular examples in existence that could be useful?

Yes. I believe that the Open Source Initiative has been working to standardize a definition of "open-source AI" which will help (https://opensource.org/blog/open-source-ai-definition-weekly-update-mar-25). In addition to this, I believe that MIT and Apache licenses are both fairly standard within the community and are compatible with one another.

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

No comment.

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

AI systems benefit from economies of scale. It seems likely to me that AI systems will mostly be deployed by very large, very visible actors (like multinational corporations) such that it would be straightforward for the government to regulate the deploying entities if it chose to. These same actors are also less likely to need others to release model weights for them, and less likely to release weights themselves. Due to these factors, I don't think that the wide availability of open foundation model weights will particularly facilitate or frustrate government action in AI regulation.

c. When, if ever, should entities deploying AI disclose to users or the general public that they are using open foundation models either with or without widely available weights?

No comment.

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 encourage standards-setting for best practices and setting metrics for risk. It should also set policies for how to manage widely-available foundation model weights in the context of military & governmental use, no matter whether that means restricting or encouraging their use. I also believe it should clarify the conditions under which foundation models can be trained, deployed, and made widely available in any form, including if there are types or sources of data that should not be trained on. I do not believe that the U.S. government should attempt in general to restrict the availability of foundation model weights.

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

Other commenters will likely write better on this. No comment.

e. What should the role of model hosting services (e.g. HuggingFace, GitHub, etc.) be in making dual-use models with open weights more or less available? Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed?

Hosting services should be expected to enforce policies on models with open weights to the extent they are selling inference services for those models. It does not seem reasonable or particularly useful to do this for platforms like GitHub (or to a certain extent, the HuggingFace Hub) that are merely *storing* the model weights, as intervening at the point of inference seems both more effective and less burdensome.

f. Should there be different standards for government as opposed to private industry when it comes to sharing model weights of open foundation models or contracting with companies who use them?

No comment.

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

No comment.

h. What insights from other countries or other societal systems are most useful to consider?

No comment.

i. Are there effective mechanisms or procedures that can be used by the government or companies to make decisions regarding an appropriate degree of availability of model weights in a dual-use foundation model or the dual-use foundation model ecosystem? Are there methods for making effective decisions about open AI deployment that balance both benefits and risks? This may include responsible capability scaling policies, preparedness frameworks, et cetera.

The government and companies should make decisions on appropriate degree of availability by considering *marginal risks from openness*, as advocated for in a paper by researchers at Stanford's Center for Research on Foundation Models (https://crfm.stanford.edu/open-fms/). That is, they should ask whether opening up the weights of their particular foundation model for broad availability would increase societal risk by intentional misuse beyond that already posed by closed foundation models and pre-existing technologies (such as web search on the internet).

j. Are there particular individuals/entities who should or should not have access to open-weight foundation models? If so, why and under what circumstances?

No comment.

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?

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

Other commenters will likely write better on this. No comment.

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 1026 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?

In my view, it is a useful metric to keep in mind, but not much policy should rest on it, and it should be open to revision. It seems likely that as hardware providers specialize their chips towards running AI foundation models, what exactly constitutes an "integer or floating-point operation" will significantly change. There is no fundamental definition for such operations, and there is already a move towards lower-precision formats and even operations on binary or trinary values. Moreover, over time it becomes possible to achieve better and better performance with the same budget of computational resources. I do think that aggregate compute quantities will continue

to be one of the most important and most reliable metric to track for governing the development of new AI systems.

c. Are there more robust risk metrics for foundation models with widely available weights that will stand the test of time? Should we look at models that fall outside of the dual-use foundation model definition?

No, there are no such robust risk metrics known. As we gain experience working with these foundation models and observing how they are used in practice, we will accumulate evidence about where the problems lie, in the same way as we now have evidence about other public safety issues. But that will take time.

9. What other issues, topics, or adjacent technological advancements should we consider when analyzing risks and benefits of dual-use foundation models with widely available model weights?

No comment.