Open Model Weights Are More Trustworthy and Less Vulnerable

Isaac Karth Jasmine Otto

March 27, 2024

1 Introduction

We are writing this public submission in response to the NTIA AI Open Model Weights RFC.[1] The authors hold PhDs in Computational Media from the University of California, Santa Cruz. We have experience in working with many forms of artificial intelligence, including the kinds of models under question. We are writing this in our capacity as concerned private citizens, not as the representative of any organization or entity.

2 Analysis

Current research in the fields of AI and of computer-supported collaborative work demonstrates that dual-use foundation models must be made trustworthy and secure. AI must not be used to shield decision-makers from responsibility for actions suggested by AI, nor from the burden of explaining their reasoning. Simultaneously, AI should not be used in capacities where it is readily exploited to subvert decision-making processes.

In many areas, the highest risk of harm is often not from the model weights, but rather what the AI is used for. It is most appropriate to involve outside stakeholders before new capabilities derived from AI models are incorporated into products, but this step is often skipped in practice, according to AI experts from industry and academia.[2] In areas where beneficial use is possible, many AI systems developed in practice have failed to incorporate the perspectives of diverse stakeholders, and as a result caused harm: "the fundamental assumptions of many educational AI systems, while motivated by logics of care, may instead reproduce structural inequities of the status quo in education"[3]

1. Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights, 89 Fed. Reg. 14059-14063 (Feb. 26, 2024)

2. Alexander Lavin et al., "Technology readiness levels for machine learning systems," *Nature Communications* 13, no. 1 (October 20, 2022): 6039, ISSN: 2041-1723, https://doi.org/10.1038/s41467-022-33128-9, https://doi.org/10.1038/s41467-022-33128-9.

3. Su Lin Blodgett and Michael Madaio, "Risks of AI Foundation Models in Education," *CoRR* abs/2110.10024 (2021), arXiv: 2110.10024, https://arxiv.org/abs/2110.10024; Michael

The wide availability of model weights does not harm either the trustworthiness of AI or its security. The availability of open model weights is in fact essential to both aims, for reasons that many diverse stakeholder groups have articulated in the research literature.

2.1 Closed models are less trustworthy

Foundation models have inherent biases which are often hard to measure,[4] can amplify human biases in the pre-training data,[5] and can contribute in inequality in areas such as healthcare[6] and education.[7] In the absence of analysis performed independently of the model owner, decision-makers will not have appropriate counter-balances on biased representations emergent in any given foundation model, nor alternative capabilities matching those of the biased model. Since foundation models will be used to support decision-making in healthcare, education, and other public spheres of concern, this issue threatens public equity.

2.1.1 AI models contain bias that is hard to detect

One significant concern for equity in AI is that models learn biases from their training data and will amplify those biases.[8] Recommender systems can be measurably unfair when making predictions for minority groups.[9] Even when image models used balanced data, they can still reflect gender biases, requiring additional effort to reduce the bias.[10]

Given a black box model, whose weights are not available to outside researchers, it can be difficult to evaluate the model's biases. Even with access to the open weights for retraining, the project struggled to retrain the model

Madaio et al., "Beyond "fairness": Structural (in) justice lenses on AI for education," in *The ethics of artificial intelligence in education* (Routledge, 2022), 203–239.

4. Beier Zhu et al., "Generalized Logit Adjustment: Calibrating Fine-tuned Models by Removing Label Bias in Foundation Models," in *Advances in Neural Information Processing Systems*, ed. A. Oh et al., vol. 36 (Curran Associates, Inc., 2023), 64663–64680, https://proceedings.neurips.cc/paper_files/paper/2023/file/cbe1fd3136e0f049bb8bc104231ccb99-Paper-Conference.pdf.

5. Tolga Bolukbasi et al., "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in *Advances in Neural Information Processing Systems*, ed. D. Lee et al., vol. 29 (Curran Associates, Inc., 2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

6. Geoff Keeling, "Algorithmic bias, generalist models, and clinical medicine," *AI and Ethics*, August 14, 2023, ISSN: 2730-5961, https://doi.org/10.1007/s43681-023-00329-x, https://doi.org/10.1007/s43681-023-00329-x.

7. Su Lin Blodgett and Michael Madaio, "Risks of AI Foundation Models in Education," *CoRR* abs/2110.10024 (2021), arXiv: 2110.10024, https://arxiv.org/abs/2110.10024.

8. Jieyu Zhao et al., *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*, 2017, arXiv: 1707.09457 [cs.AI].

9. Sirui Yao and Bert Huang, "Beyond Parity: Fairness Objectives for Collaborative Filtering," in *Advances in Neural Information Processing Systems*, ed. I. Guyon et al., vol. 30 (Curran Associates, Inc., 2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf.

10. Tianlu Wang et al., "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 5309–5318, https://doi.org/10.1109/ICCV.2019.00541.

to recognize the cultural terms. With a closed model, minority groups must petition the centralized model owner to perform the training, making equity of representation dependent on the largess of the corporation, who is often not in a position to evaluate the equity of representation. With an open model independent groups can retrain their own models, furthering the equity of access to the technology.

One approach proposed for evaluating biases in black box models has been to query the model itself. However, the model's response is not a reliable 'explanation' of its own behavior. Because model training selects for responses that are believable, without regard to truth in any particular context, conversational models are unbound by intuitive conversational assumptions of joint accountability.[11] Their responses to queries phrased as questions about themselves resemble answers to that question, but are not actual *answers*. Therefore, narrow strategies of evaluation such as "only evaluating the plausibility of explanations [...] may increase trust in AI systems without guaranteeing their safety."[12]

2.1.2 Independent audits of models are desirable and necessary

Open access to the model weights is important for accurately auditing them.[13] A foundation model which cannot reliably be accessed by independent investigators is not trustworthy. If the model weights are not made available, then investigators cannot reproduce the model using their own computing resources. Instead they would rely upon the good-faith involvement of the model owner, which cannot be assumed.

As with any other statistical technique, classifications made by foundation models may or may not correlate with ground-truth observations about the world. This epistemic uncertainty has consequences for decision-makers using AI support tools, as a Microsoft Research team found in their survey of 60 papers about over-reliance on AI[14], defined in terms of *user agreement with incorrect AI recommendations*. Because "it is not always obvious when users over-rely on AI", proactive analysis of AI models by independent investigators should be cultivated, so that alternative versions of the capability can be discerned and produced.

---

11. Mark Dingemanse and N.J. Enfield, "Interactive repair and the foundations of language," *Trends in Cognitive Sciences* 28, no. 1 (2024): 30–42, ISSN: 1364-6613, https://doi.org/https://doi.org/10.1016/j.tics.2023.09.003, https://www.sciencedirect.com/science/article/pii/S1364661323002504.

12. Miles Turpin et al., "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting," in *Advances in Neural Information Processing Systems*, ed. A. Oh et al., vol. 36 (Curran Associates, Inc., 2023), 74952–74965, https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

13. Stephen Casper et al., *Black-Box Access is Insufficient for Rigorous AI Audits*, 2024, arXiv: 2401.14446 [cs.CY].

14. Samir Passi and Mihaela Vorvoreanu, "Overreliance on AI: Literature review" (https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf, 2022).

2.1.3 Open access to model weights increases equity

Foundation models are prone to not representing minority concepts unless they are amended by the minority community. For example, the Reservation of the Pala Band of Mission Indians successfully audited an AI model and realized it was blind to their culture. A joint project between the San Diego Supercomputer Center and the tribal band attempted to use transfer learning to re-train an image model to better represent specific sacred, traditional performing arts.[15]

Access to open weights model weights allows groups affected by algorithmic inequality to audit the model's biases, create datasets, and retrain the models to better reflect the needs of their community.

2.2 Closed models are more vulnerable

We argue that closed foundation models are not only less trustworthy, but also less secure than comparable models with open weights.

2.2.1 Clones of closed models become rapidly available

New models have seen rapid release. Just hitting the highlights, OpenAI's ChatGPT was released in November 2022, based on GPT 3.5. Meta's Llama 1 followed in February 2023, as a foundation model.[16] Stanford's Alpaca replicated ChatGPT's instruction tuning on March 13 2023.[17] Anthropic released the first version of Claude on March 14, 2023.[18] Abu Dhabi's Technology Innovation Institute's Falcon 40B was made available for open use in May.[19] MosaicML's MPT-7B was also released in May.[20] Meta released Llama 2 in July 2023.[21] The French foundation model Mistral 7B v0.1 was released in the Fall of 2023.[22] Alibaba Cloud's Tongyi Qianwen was revealed in April 2023 and released to the public as an open model in September 2023.[23] Many of these models outperform

15. Kimberly Mann Bruch, "AI Applications to Illustrate Native American Arts: Birdsongs: Using Transfer Learning to Augment Image Generation Models" (Presentation, Creative AI Session 1, December 12, 2023), NeurIPS 2023, New Orleans, Louisiana, USA., https://neurips.cc/virtual/2023/event/81782.

16. Hugo Touvron et al., *LLaMA: Open and Efficient Foundation Language Models*, 2023, arXiv: 2302.13971 [cs.CL].

17. Rohan Taori et al., *Stanford Alpaca: An Instruction-following LLaMA model*, https://github.com/tatsu-lab/stanford_alpaca, 2023.

18. Anthropic, *Introducing Claude*, March 2023, https://www.anthropic.com/news/introducing-claude.

19. Lisa Barrington, *Abu Dhabi makes its Falcon 40B AI model open source*, May 2023, https://www.reuters.com/technology/abu-dhabi-makes-its-falcon-40b-ai-model-open-source-2023-05-25/.

20. The MosaicML NLP Team, *Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs*, May 2023, https://www.databricks.com/blog/mpt-7b.

21. Meta, *Meta and Microsoft Introduce the Next Generation of Llama*, July 2023, https://about.fb.com/news/2023/07/llama-2/.

22. Albert Q. Jiang et al., *Mistral 7B*, 2023, arXiv: 2310.06825 [cs.CL].

23. Casey Hall, *Alibaba opens AI model Tongyi Qianwen to the public*, 2023, https://www.reuters.com/business/retail-consumer/alibaba-opens-ai-model-tongyi-qianwen-public-2023-09-13/.

GPT 3.5 on at least some metrics, such as context length or resource use. All of them are open, to a greater or lesser extent.

Despite OpenAI keeping the closed weights of ChatGPT and GPT 3.5 locked behind an API, this timeline demonstrates that it took very little time for its capabilities to be replicated. While ChatGPT has some functionality that these other instruction models do not—such as presenting the user interaction as an ongoing conversation—most of that functionality is part of the apparatus around the model weights, rather than the weights themselves.

Knowledge distillation means API isn't much of a moat either You might naively assume that keeping a model closed off behind an API would be safer than releasing the model weights, similar to a store locking the expensive electronics behind glass doors. However, this is ineffective in practice because of a technique called knowledge distillation.[24] Knowledge distillation (and knowledge transfer) means that you can take the output of a "teacher" model and train a "student" model to have comparable capabilities.[25] Access to the model weights is helpful, but so-called *black-box knowledge distillation*[26] involves training solely on the output of the teacher model.[27] The API itself provides enough data. Locking the model weights in the metaphorical glass case of an API isn't effective in the long term.

Even access to the API may be unnecessary: for ChatGPT in particular, enough of the outputs of the model have been posted publicly on the internet to make data contamination by ChatGPT something that requires active effort to *prevent*. Controlled amounts of synthetic training data enhances performance, but after consuming too much self-generated text, performance collapses,[28] so knowing how much data is model-generated can be important. Similar to pre-atomic steel being sought out for use in precision medical devices, pre-ChatGPT data is valuable because it doesn't need to be filtered for AI output. However, if an attacker wanted to train a model that closely matched ChatGPT's performance, the publicly available text may be enough to clone many of the model's responses without ever needing access to the API.

Therefore, limiting access is only partially effective at stopping bad actors.

24. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, *Distilling the Knowledge in a Neural Network*, 2015, arXiv: 1503.02531 [stat.ML].

25. Jang Hyun Cho and Bharath Hariharan, "On the Efficacy of Knowledge Distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2019).

26. Yuxian Gu et al., *Knowledge Distillation of Large Language Models*, 2024, arXiv: 2306.08543 [cs.CL].

27. Rohan Taori et al., *Stanford Alpaca: An Instruction-following LLaMA model*, https://github.com/tatsu-lab/stanford_alpaca, 2023; Wei-Lin Chiang et al., *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*, March 2023, https://lmsys.org/blog/2023-03-30-vicuna/; Minghao Wu et al., *LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions*, 2024, arXiv: 2304.14402 [cs.CL]; Baolin Peng et al., *Instruction Tuning with GPT-4*, 2023, arXiv: 2304.03277 [cs.CL].

28. Martin Briesch, Dominik Sobania, and Franz Rothlauf, *Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop*, 2023, arXiv: 2311.16822 [cs.LG].

However, having access to the model weights provides many benefits to positive innovation: while ChatGPT demonstrated that instruction models were possible, it took an open model being available for research into them to really get going with the release of Alpaca.[29] Further research into instruction model training–such as LIMA, which used a 65-billion parameter version of Llama[30]—relied on the availability of open weights. Future research into making interacting with instruction models safer and more effective depends on open access to models that are large enough for effective evaluation.

2.2.2 Security through obscurity is ineffective

The practice of keeping secrets in hopes of making a system more secure is sometimes termed "security through obscurity." It is generally ineffective on its own, offering, at best, temporary security while limiting the visibility of actual exploits. For other technologies, guidelines frequently recommend avoiding relying on secrecy. As an example, NIST guidelines for securing servers emphasizes open design: "System security should not depend on the secrecy of the implementation or its components."[31] Likewise the Common Weakness Enumeration lists "Reliance on Security Through Obscurity" as common weakness "CWE-656": "The product uses a protection mechanism whose strength depends heavily on its obscurity, such that knowledge of its algorithms or key data is sufficient to defeat the mechanism."[32]

While secrecy can slow down some exploits, it can also make a system more vulnerable if the attackers discover weaknesses that are unknown to the original developers. This is a very common occurrence in computing in general. Given the very limited extent of our current understanding of the fundamental properties of these models, it is very difficult to secure them behind an API when we often don't know what a vulnerability looks like. Take the example of adversarial attacks, which are just one category of attacks on a model. If we hope to do something to mitigate adversarial attacks on important systems, it is vital that we have enough large open models for researchers to experiment on defending against those attacks, which makes it much easier for the development and dissemination of mitigation and prevention strategies. Making future models safer requires widespread open access now.

29. Rohan Taori et al., *Stanford Alpaca: An Instruction-following LLaMA model*, https://github.com/tatsu-lab/stanford_alpaca, 2023.

30. Chunting Zhou et al., *LIMA: Less Is More for Alignment*, 2023, arXiv: 2305.11206 [cs.CL].

31. K A Scarfone, W Jansen, and M Tracy, *Guide to general server security* (2008), https://doi.org/10.6028/nist.sp.800-123, http://dx.doi.org/10.6028/NIST.SP.800-123.

32. *CVE-656*, CVE-ID CVE-656, January 2008, accessed February 29, 2024, https://cwe.mitre.org/data/definitions/656.html.

Is a regulation worth implementing if it only buys you a couple of months? Regulating the distribution of model weights is, therefore, strictly limited in terms of how long any benefits will last. Merely knowing that something can be done has often been enough to replicate it (as in the case of Stanford's Alpaca). Training a foundation model from scratch does take a significant amount of resources, but the past year has shown us that the amount of time that any given model is state-of-the-art is limited.

The models at the performance frontier (i.e., GPT-4) maintain their edge a little bit longer. Since it is impossible to examine the inner workings of GPT-4, some of GPT-4's market edge is likely to be first mover advantage and branding, rather than the actual performance metrics.[33] Despite the GPT-4 technical report deliberately omitting all information about "the architecture (including model size), hardware, training compute, dataset construction, training method, or similar"[34] other models have been able to approach GPT-4's performance, so even extreme secrecy has limited effect at withholding performance from bad actors while making it difficult for good actors to learn from OpenAI's safety measures.

This speed may change in the future, but, for the question of the timeframe of equivalent-models becoming available (question 1.a) it's reasonable to conclude that for the foreseeable near-term future, any benefit of restricting the distribution of model weights will have a strict time-limit. Commercial restrictions on the availability of model weights has already encouraged nation-state-level investment into training new foundation models. Further restrictions will inevitably spur the creation of more non-US models, eroding any lead by the United States in this field. Making it difficult to access open model weights will work against the the executive order's goal to "attract the world's AI talent to our shores" and will negatively affect the goal of ensuring that "the companies and technologies of the future are made in America." (E.O. section 2.b) while providing relatively limited benefits.

2.2.3 Increased access does not imply greater risk

The "Gradient of Access" proposes that AI models are initially fully closed; as documentation, API access, and downloadable access is released piece by piece, their openness increases[35], and therefore so does exposure to various categories of risk, while reducing others. While keeping a model secret might temporarily prevent misuse of that particular model, it prevents the wider research community from fully evaluating its vulnerabilities (Sec. 2.2.2, page 6). This problem is compounded as successive generations of models increase their

33. GPT-4 is known to favor its own output when asked to evaluate text, which is called "self-enhancement bias" (Lianmin Zheng et al., *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, 2023, arXiv: 2306.05685 [cs.CL]). This makes it difficult to accurately compare its performance with other models when using it as part of an automatic LLM-based evaluation.

34. OpenAI et al., *GPT-4 Technical Report*, 2024, arXiv: 2303.08774 [cs.CL].

35. Irene Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, 2023, arXiv: 2302.04844 [cs.CY].

capabilities: because fewer researchers were able to test the earlier model, the newer model combines more potential harm with existing but undetected vulnerabilities. Limiting the access to the model does not fully prevent exploitation: empirical testing has revealed ChatGPT jailbreak prompts that have gone unpatched for over 100 days.[36]

Even individual model risk can't be ranked on a linear scale. Some methods of limiting access to model weights appear to be safer than they actually are, depending on the threat being defended against. Local model weights are often more expensive to operate than API access, due to economies-of-scale and loss-leader pricing. Self-hosted local model weights also require more technical knowledge to operate. While there have been many efforts to make running models easier, using a model in production often takes expensive skilled labor to implement and maintain. For individual-level harms, it is far cheaper and more efficient for a malicious user to jailbreak a hosted service rather than acquire the skills and hardware to run the model themselves. ChatGPT had over 100 million monthly active users in January 2023;[37] as of March 2023 the open text generation model with the most downloads on huggingface.co is GPT-2 at 7.5M downloads.[38] There are at least two orders of magnitude more people using the ChatGPT service, versus the number of people who have *ever* downloaded the most popular open model. For nation-state actors, the worldwide shortage of GPUs means that using US companies' hardware via APIs is an attractive attack vector.

Further, there's a natural limit on the size of models: bigger models require much more expensive hardware. While quantization has made it easier to run large models on consumer hardware, the larger models still require data-center-class hardware. 10-billion parameter models can run on single consumer GPUs, 100-billion parameter models require a larger investment, one that is within the reach of small research labs but out of the reach of a casual individual. Limiting access to open models that are already limited by hardware constraints would harm research without causing much reduction of risk.

Larger models may not have suddenly greater capabilities There has been concern that larger model exhibit emergent effects—that is, bigger models can unexpectedly do things that smaller models can't, implying that performance will exhibit significant jumps as models scale up. Recent research has cast doubt on this view, suggesting that many of the so-called emergent effects are due to how the model was measured, rather than the actual abilities of the model.[39] Sudden jumps in model capability may just be a side-effect of the

36. Xinyue Shen et al., *"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models*, 2023, arXiv: 2308.03825 [cs.CR].

37. Krystal Hu, *ChatGPT sets record for fastest-growing user base - analyst note*, February 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

38. huggingface.co, *Huggingface Models, sorted by most downloads*, March 2023, https://huggingface.co/models?pipeline_tag=text-generation&sort=downloads.

39. Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, *Are Emergent Abilities of Large Language Models a Mirage?*, 2023, arXiv: 2304.15004 [cs.AI].

measurements, rather than a suddenly emergent new capability.

Therefore, concerns that there is a size where a large model will spontaneously reveal unexpected capabilities may be overblown. Unexpected capabilities will, in this case, come from active research efforts rather than uncontrolled model growth. The more researchers who have access to open models, the better chance we have of anticipating unexpected capabilities, and therefore the lower the overall risk will be.

3 Conclusion

Open access to model weights is important for building safe and trustworthy AI. Closed models are inherently less able to be evaluated, making them brittle: their safety relies on obscuring access. While closed models could be made somewhat more trustworthy by involving outside stakeholders in evaluating models, this is often skipped in practice.

Because it can be hard to detect biases and weaknesses in models, it is vital that researchers have access to open weights, so that they can independently audit them. Open models allows groups affected by model inequality to independently assess and retrain the models that affect them.

Keeping a model closed models makes it more vulnerable to certain attacks. Security through obscurity is brittle. Clones of closed models have been created relatively rapidly, and are available from many different countries. Knowledge distillation means that the mere availability of output generated by a model can be enough to create a functional-enough clone.

In other areas of computation, relying on secrecy as the main defense is regarded as a vulnerability in itself: secrecy means that friendly researchers can't assist in detecting and patching vulnerabilities, while still leaving you exposed to attackers who discover the exploits. Limiting access to the weights for the purposes of AI safety is a short-term solution with a high risk of leaving critical vulnerabilities undetected.

Limiting software distribution is difficult and ineffective, and too many restrictions on the distribution of open weights will damage the future development of AI in the United States, risking the country's current leadership in innovation.

The "Gradient of Access" model should not be understood as a linear scale of increasing risk: too little scrutiny can lead to critical vulnerabilities being overlooked, even when a model is never released as an open model. Further, the risk from any one set of open model weights being available must be weighed against the need to have the current generation of models audited for vulnerabilities so they can be caught and addressed at this stage, rather than remaining hidden until they inevitably surface in a later (and more capable) generation of models.

While there has been some concern about larger models exhibiting emergent effects, recent research suggests that model capabilities do not have sudden jumps in their capabilities as they scale up: a bigger model is more capable, but

in a predictable way.

Overall, access to open model weights is important for continuing research into making the models safer, more trustworthy, and more equitable. Closed models may seem, at first glance, to be safer because they lock down access to the model, but this security is brittle and hides vulnerabilities that could have been discovered and fixed in an open model.

3.1 RFC questions

Some of the questions we endeavored to answer in this public submission are summarized below.

1.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?

Closely-equivalent models have become available–often within a few months—of currently-closed AI systems; keeping model weights secret doesn't buy much time but does suppress opportunities for innovation, including innovation in safety guardrails (Sec. 2.2.1, page 4).

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

Closed model weights might seem, at first glance, to be safer than open model weights, but their security can be deceptively brittle (Sec. 2.2.2, page 6). Making model weights open allows for independent auditing (Sec. 2.1.2, page 3) and gives vulnerable groups the ability to examine and retrain the models to reflect their needs 2.1.3, page 4).

3.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?

Open dual-use foundation models make it possible for widespread research participation by individuals, research labs, and small businesses (Sec. 2.1.2, page 3). Too-stringent restrictions on the distribution of model weights will encourage AI talent and innovation to leave America, but any safety benefits will be short-lived (Sec. 2.2.2, page 7).

3.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?

Making model weights widely available improves safety, security, and trustworthiness of AI by making it possible for independent researchers to audit and revise the models. (Sec. 2.1.2, page 3). Closed models, in contrast, are brittle: they still have the same vulnerabilities, but it is much harder for the vulnerabilities to be discovered before they are exploited by malicious actors (Sec. 2.2.2, page 6). Security through obscurity is ineffective, and despite the "gradient of

access" we should be careful about assuming that the secrecy of closed models will protect us from malicious attacks more effectively than openly audited models (Sec. 2.2.3, page 7)

3.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.)?

Biases in foundation models can contribute to inequality (Sec. 2.1, page 2). Open model weights advance equity by allowing researchers to independently evaluate the models (Sec. 2.1.2, page 3), and by empowering minority communities to have the option of creating their own datasets and models (Sec. 2.1.3, page 4).

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

The executive order defines the dual-use foundation models as containing "tens of billions of parameters" (E.O. Section 3.k.), which may be unnecessarily low given hardware limitations; doubts about the existence of emergent effects means that concerns about the model size unexpectedly surpassing a critical threshold may be misguided (Sec. 2.2.3, page 8).