THE
FUTURE
SOCIETY

Response to the NTIA Request for Comment: Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

Prepared by The Future Society (TFS)

March 2024

Response to Openness in AI Request for Comment
Docket number: NTIA–2023–0009

We are grateful for this opportunity to submit a response on *Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights*.

The Future Society (TFS) is a U.S. 501(c)(3) independent nonprofit organization with a mission to align artificial intelligence through better governance. Our policy research aims to ensure safety and adherence to fundamental values in the development and deployment of advanced AI systems.

In December 2023, we hosted interactive, small-group roundtable discussions on 'Navigating AI Deployment Responsibly: Open-Source, Fully-closed, and the Gradient in Between' with over 100 leading AI policy experts at our fifth edition of The Athens Roundtable on AI and the Rule of Law in Washington, D.C. The key takeaways from these discussions have informed this response and may be found in our summary report "Towards Effective Governance of Foundation Models and Generative AI."

Thank you for your consideration of our comments. We may be contacted at yolanda.lannquist@thefuturesociety.org or info@thefuturesociety.org.

We begin with responses to questions 7.i. and 8 and then proceed in numerical order.

> *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?*
>
> *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.*

Pre-release evaluations allow for the detection and mitigation of potential risks associated with dual-use foundation models, while still allowing for the realization of their benefits.

As models become capable of performing and assisting with a broader range of functions, pre-release evaluations (i.e., evaluations conducted before a model is made accessible beyond its developers or its testing environment), will play a more necessary role in identifying and addressing risks before models can cause significant damage, irrespective of their release method.

Pre-release evaluations can be used to identify and remedy deficiencies and potential risks—including privacy breaches, biased outputs, and security vulnerabilities—and to determine which steps, if any, are necessary to remedy identified issues, and which release method(s) are appropriate for the model in question.

The term "pre-release evaluations," as with many terms used in AI measurement, lacks a universally accepted definition and might be used to refer to any combination of the following:

- **Benchmarks**: Standardized datasets used to assess and compare the performance of different models on specific tasks or capabilities.[1]
- **Model evaluations**: Standardized protocols conducted to measure model performance, robustness, fairness, and other key attributes, such as testing a model's ability to acquire resources.[2]
- **Red-teaming tests**: Adversarial testing methods that attempt to find vulnerabilities, biases, or unintended behaviors in a model or its cybersecurity environment.[3]

[1] See, for example, Holistic Evaluation of Language Models (HELM).
[2] See, for example, New report: Evaluating Language-Model Agents on Realistic Autonomous Tasks.
[3] See, for example, What Does AI Red-Teaming Actually Mean? | Center for Security and Emerging Technology.

Each approach has its own benefits and drawbacks in terms of resource requirements, validity, robustness, and scalability.

Each approach provides a distinct and necessary benefit: Benchmarks provide standardized performance comparisons across models; model evaluations allow for more in-depth assessments of a model's attributes and behaviors in an operating environment; and red-teaming tests enable the identification of hard-to-anticipate vulnerabilities and risks. As such, in the interest of exercising precaution, **pre-release evaluations should incorporate a combination of each approach**.

At the same time, unduly onerous pre-release evaluation requirements would function as a bottleneck for market entrants or otherwise prevent the benefits of foundation models, including open models, from being realized. With this in mind, pre-release evaluations should only be required by a **limited set of models** (i.e., those considered to be "dual-use foundation models"). To counteract potential market concentration, the government could **provide guidance and services to SMEs** developing models that necessitate pre-release evaluations. Additionally, to ensure rigor and objectivity of the pre-release evaluation, at least some of the components, such as model evaluations and red-teaming, should be conducted by **independent, qualified third parties**. Red-team testers should be incentivized to continuously use creative techniques to elicit undesirable behavior and upper-bound capabilities through the provision of "bug bounties."[4] Joint and several liability regimes could be explored as a means to incentivize both developers and third parties to perform exhaustive evaluations.

In conjunction with model pre-release evaluations, the government should establish **risk categories (i.e., designations of "high-risk" or "unacceptable-risk"), thresholds, and risk-mitigation measures** that correspond to evaluation outcomes. Entities developing, providing, or deploying models designated as "high-risk" should be required to satisfy enhanced oversight requirements, while the possession, use, and proliferation of models designated as "unacceptable-risk" should be restricted.

Due to the evolving nature of AI measurement science, and of models of themselves, **pre-release evaluations must adapt as more robustly valid methods are demonstrated**.

Finally, model pre-release evaluations are just one method of reducing risk presented by dual-use foundation models; **they should complement—and do not eliminate the need for—other risk assessment or mitigation strategies**, including testing at the AI system- or application-level.

[4] Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework.

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?

Recognizing that we lack sufficient data to make informed policy decisions, it is necessary to undertake a precautionary approach that seeks to mitigate potential risks while still fostering a diverse and competitive AI ecosystem.

We expect that much of the evidence cited in responses to this RFC will have been produced within the past 12 months.

The current state of AI measurement science is insufficient to fully understand the long-term benefits and risks of open foundation models. Steps should immediately be taken to advance measurement science—e.g., investments should be made in the development of standardized evaluation and risk-assessment frameworks, and information-gathering measures, such as reporting requirements, post-market monitoring, and incident reporting,[5] should be institutionalized.

At the same time, we are concerned that overly restrictive policies could lead to market concentration, hindering competition and innovation in both industry and academia. A lack of competition in the AI market can have far-reaching knock-on consequences, including potentially stifling efforts to improve transparency, safety, and accountability in the industry. This, in turn, can impair our ability to monitor and mitigate the risks associated with dual-use foundation models and to develop evidence-based policymaking.

---

1. How should 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?

In March 2023, the weights for Meta's LLaMa model, which were intended to only be accessed by approved researchers, were leaked online.[6] At that time, the model was comparable to contemporaneous closed-source models, such as GPT-3, Chinchilla, and PaLM models.[7] In the time since, there have been several instances of companies deliberately releasing weights for models that are comparable to contemporaneous closed-source models, such as Meta's release

[5] An Argument for Hybrid AI Incident Reporting | Center for Security and Emerging Technology
[6] Meta's powerful AI language model has leaked online — what happens now? - The Verge.
[7] [2302.13971] LLaMA: Open and Efficient Foundation Language Models

of LLaMa-2 in July 2023,[8] Mistral's release of Mixtral 8x7B in December 2023,[9] and Databricks's DBRX in March 2024.[10]

Another route that weights similar to (or rather, *matching*) those of currently-closed models could become widely accessible is through the exfiltration of current providers' models. An interim report published by RAND researchers in 2023 identified approximately ~40 distinct attack vectors that could compromise providers' security.[11] We are not aware of a public, large-scale assessment of providers' cybersecurity systems, but one small-scale study by an independent researcher in February 2024 found that 11 out of 16 leading AI companies evaluated had fallen short of voluntary commitments to cybersecurity vulnerability reporting—a small yet nontrivial component of cybersecurity infrastructure.[12]

> *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?*

Access to model weights allows, through fine-tuning, for the removal or circumvention of guardrails that might have been incorporated in the production of the model.[13] This could allow models to more easily be used in a malicious manner, including in spear-phishing scams[14] deepfake pornography,[15] voice cloning scams,[16] child sexual abuse material[17], and in techniques to "jailbreak" closed-source foundation models.[18] Outside of the US, these effects may be more acute in countries with fragile democratic institutions, weak cyber resilience, susceptibility to social unrest, or lacking legal or regulatory safeguards or the capacity to enforce them.

---

[8] Llama 2
[9] Mixtral of experts | Mistral AI | Frontier AI in your hands
[10] Databricks Launches DBRX, A New Standard for Efficient Open Source Models
[11] Securing Artificial Intelligence Model Weights: Interim Report | RAND
[12] How are AI companies doing with their voluntary commitments on vulnerability reporting?
[13] A recent paper by Stanford University acknowledges that "for open-source models safety filters can simply be removed by deleting a few lines of code." Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models. Also see BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B
[14] WormGPT - The Generative AI Tool Cybercriminals Are Using to Launch BEC Attacks | SlashNext
[15] "Because the model weights were shared publicly, users easily circumvented filters that Stability AI implemented to prevent the generation of not safe for work (NSFW) imagery. As a result, AI-generated pornography based on Stable Diffusion offshoots quickly spread across the internet, including images resembling real people generated without their consent." On the Societal Impacts of Open Foundation Models.
[16] Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning.
[17] Generative ML and CSAM: Implications and Mitigations
[18] LLM Attacks

Recent research has put forward a framework for assessing the marginal risk of open vs. closed models.[19] While it argues that many of the studies that it reviewed failed to demonstrate a significant marginal risk for open models, in many cases due to lack of evidence of societal impacts, that does not invalidate the possibility that these risks might arise in the future as open foundation models are increasingly deployed.

Finally, models released with widely available weights cannot credibly be monitored, updated, or retracted. Once these models are released to the public, developers can easily and inexpensively modify or distribute them. Thoroughly assessing the risks of release is therefore crucial; however, industry actors currently lack a standardized, transparent approach to this assessment.

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

The wide availability of model weights used in audio, image, and video generation models heightens the risk of the production of "deepfakes" that imitate a person's identity without their permission. This is because availability of model weights could allow actors to circumvent features in the underlying base model or the overarching system designed to refuse requests to generate explicit images or images containing public figures.[20] According to the UN Interregional Crime and Justice Research Institute (UNICRI), deepfakes, often combined with social media and messaging applications, can reach millions of people in very short periods of time, and as such, present "considerable potential" for a range of malicious and criminal purposes, including but not limited to destroying the credibility of individuals or institutions, harassing individuals, perpetrating extortion or fraud, falsifying identities, inciting acts of violence towards minority groups, or supporting narratives of extremist or terrorist groups.[21]

> *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?*

Making model weights widely available allows for more in-depth and reproducible experimentation, transparency, evaluations, audits, and improvements that can bolster

[19] On the Societal Impacts of Open Foundation Models
[20] "Because the model weights were shared publicly, users easily circumvented filters that Stability AI implemented to prevent the generation of not safe for work (NSFW) imagery. As a result, AI-generated pornography based on Stable Diffusion offshoots quickly spread across the internet, including images resembling real people generated without their consent." On the Societal Impacts of Open Foundation Models. Also see Suspect Charged After Allegedly Using AI To Create Images Of Child Sexual Abuse
[21] UNICRI, Malicious Uses and Abuses of Artificial Intelligence. Page 52, Case Study: A deep dive into deepfakes.

robustness and trustworthiness. However, these tasks do require knowledge and resources that may limit the number of contributors.[22]

> *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?*

Several existing U.S. laws and regulations related to intellectual property (e.g., Copyright Act, Patent Act, Trade Secrets Act), data privacy (e.g., HIPAA, COPPA, CCPA), and national security (e.g., ECRA, EAR, ITAR) may indirectly affect the release of certain AI models or their components, including open source models, by imposing restrictions on their use, reproduction, distribution, export, or foreign investment; however, the application of these laws to AI models is still evolving, open to interpretation, and new regulations may emerge as AI technology advances.

To prevent wide availability of access to a foundation model's weights, or limit their end use, policymakers might establish coherent liability frameworks that cover the entire lifecycle of dual-use foundation AI models, including legal liability for open source providers. This could involve establishing legal liability for open-source providers based on concepts such as "reasonably foreseeable misuse," as proposed in the draft EU Cyber Resilience Act.

Additionally, existing regulatory entities could expand their mandates to encourage transparency, red-teaming, and promote competition among closed model developers, reducing the reliance on open-source models as a means to achieve these goals. For instance, the FTC and the DOJ could intensify their efforts to scrutinize anticompetitive practices, enforce antitrust laws, and foster innovation in the AI industry. Closed AI models can be subject to transparency requirements and independent red-teaming exercises to identify dangerous capabilities, vulnerabilities, or other emergent properties of foundation models (on the base models since guardrails can be removed). Red-teaming exercises must adhere to standards that are at least as rigorous as forthcoming NIST guidelines for AI evaluation and red-teaming.

---

[22] Widder, D. G., West, S. and M. Whittaker. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. *"We find that 'open' AI can, in its more maximal instantiations, provide transparency, reusability, and extensibility that can enable third parties to deploy and build on top of powerful off-the-shelf AI models. These maximalist forms of 'open' AI can also allow some forms of auditing and oversight. But even the most open of 'open' AI systems do not, on their own, ensure democratic access to or meaningful competition in AI, nor does openness alone solve the problem of oversight and scrutiny."*

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

Models released with widely available weights cannot credibly be monitored, updated, or shut down. Once these models are released to the public, open-source developers can easily and inexpensively modify, distribute, and enhance them. This makes it difficult to conduct oversight, incident reporting, and implement regulatory requirements post-release. Furthermore, it makes it difficult, if not impossible, to trace and attribute instances of misuse and seek redress.[23]

Second, the wide availability of open foundation model weights can undermine policies intended to bolster U.S. technological leadership by enabling the transfer of potentially millions of dollars worth of U.S. R&D to foreign firms or state actors, allowing foundation model developers from other nations to easily replicate (and potentially improve upon) U.S. advancements in AI. This was exemplified when 01.AI, a prominent Chinese foundation model developer, was found to have partly copied the transformer architecture from Meta's open-source Llama-2 model.[24]

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

To reduce the compliance burden of developers without sacrificing safety, it would be useful to maintain consistency with the requirements imposed on developers under the EU AI Act. Under the EU AI Act, all providers of general-purpose AI (i.e. foundation) models that present a "systemic risk"—open or closed—must also conduct state-of-the-art tests and model evaluations, report serious incidents, assess and mitigate risks, meet cybersecurity requirements and provide information about their models' energy consumption. General-purpose AI models are considered to pose "systemic risk" under the AI Act when the cumulative amount of compute used to produce them exceeds 10^{25} floating point operations (FLOP).[25]

[23] Towards Effective Governance of Foundation Models and Generative AI - The Future Society
[24] Chinese Startup 01.AI Is Winning the Open Source AI Race | WIRED; "China's Rush to Dominate A.I. Comes With a Twist: It Depends on U.S. Technology."
[25] European Parliament, Artificial Intelligence Act, (March 2024): https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf