Department of Commerce
National Telecommunications and Information Administration
Docket No. 240216-0052
RIN 0660-XC06
Request for Comment on Dual Use Foundation Artificial Intelligence Models With
Widely Available Model Weights

March 27, 2024

1. Introduction. Foundation models present tremendous benefits and risks to society as central artifacts in the AI ecosystem. In addressing dual use foundation models with widely available weights, the National Telecommunications and Information Administration (NTIA) should consider the *marginal risk* of open foundation models,[1] defined as the extent to which they increase risk relative to closed foundation models or preexisting technologies like search engines. Open foundation models also have a number of unique benefits: They can catalyze innovation, increase transparency, enable science, and combat the concentration of power. The Commerce Department should seek to amplify these benefits and further assess the extent of marginal risks. For further details, please see our recent paper on the societal impact of open foundation models.[2]

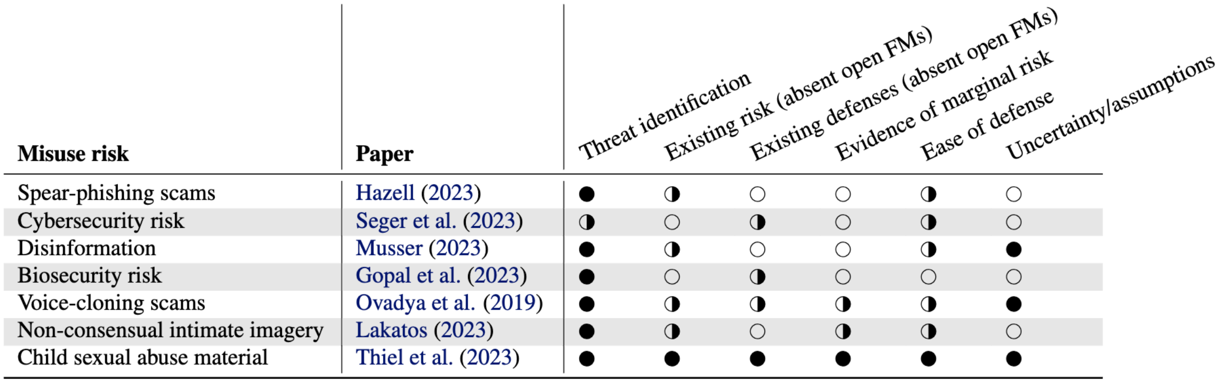
Contextualizing the risks posed by open foundation models

2. The federal government should prioritize understanding of marginal risk. The risks of open foundation models do not exist in a vacuum. To properly assess the risks of open foundation models, and whether regulations should single out open foundation models, the federal government should directly compare the risk profile to those of closed foundation models and existing technologies. In its report, the NTIA should foreground the *marginal risk* of open foundation models by directing government agencies to conduct marginal risk assessments, fund marginal risk assessment research, and incorporate marginal risk assessment into procurement processes.

3. Implement a risk assessment framework for open foundation models. In our recent paper, "On the Societal Impact of Open Foundation Models," we develop a framework for assessing marginal risk for open foundation models. Assessing the risk of an open foundation model, whether as an academic researcher or a government agency, requires identifying (i) the threat model, (ii) the background conditions (existing attacks and defenses), (iii) evidence of the marginal risk, (iv) resilience to marginal risk, and (v) uncertainty/assumptions. In reviewing the scientific literature, we examine risks related to spear-phishing scams, cybersecurity, disinformation, biosecurity, voice-cloning, non-consensual intimate imagery (NCII), and child sexual abuse material (CSAM).

[1] We define open foundation models as those with widely available weights in line with the request for comment on Dual Use Foundation Artificial Intelligence Models With Widely Available Weights.
[2] Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," *arXiv*, February 27, 2024, https://arxiv.org/abs/2403.07918.

We found that for six of the seven areas, the studies we analyzed do not provide persuasive evidence for the marginal risk of open foundation models: They do not consider steps in the framework such as what is possible with existing technologies or how defenses will adapt to marginal risks. However, for CSAM-related risks, Thiel et al. (2023)[3] conducted a complete analysis that shows marginal risks from open foundation models that are not satisfactorily addressed.[4] To provide guidance, we conducted preliminary marginal risk assessments for automated cybersecurity vulnerability detection and NCII, and we found that current marginal risk of open foundation models is low for automated vulnerability detection (due in part to the efficacy of AI for defense), whereas marginal risk of open models for NCII is considerable.



Misuse risk	Paper	Threat identification	Existing risk (absent open FMs)	Existing defenses (absent open FMs)	Evidence of marginal risk	Ease of defense	Uncertainty/assumptions
Spear-phishing scams	Hazell (2023)	●	◑	○	○	◑	○
Cybersecurity risk	Seger et al. (2023)	◑	○	◑	○	◑	○
Disinformation	Musser (2023)	●	◑	○	○	◑	●
Biosecurity risk	Gopal et al. (2023)	●	○	◑	○	○	○
Voice-cloning scams	Ovadya et al. (2019)	●	◑	◑	◑	◑	●
Non-consensual intimate imagery	Lakatos (2023)	●	◑	○	◑	◑	○
Child sexual abuse material	Thiel et al. (2023)	●	●	●	●	●	●

Table 1. Misuse analyses of open foundation models assessed under our risk framework (§5.1). ● indicates the step of our framework is clearly addressed; ◑ indicates partial completion; ○ indicates the step is absent in the misuse analysis. Incomplete assessments do not indicate that the analysis in prior studies is flawed, only that these studies, on their own, do not show an increased marginal societal risk stemming from open foundation models. We provide more details for our assessment of each row in Appendix B.

4. Decisive policy recommendations require additional evidence of marginal risk. Given this analysis of the state of marginal risk assessment, there is limited available evidence at present for the NTIA to make a decisive recommendation on the risks of open foundation models. Claims that open foundation models pose risks that are unique, unprecedented, or especially difficult to mitigate must show how open models differ from search engines, Wikipedia, open-source software, closed models, and other potential threat vectors. Based on a review of threat vectors, some risk dimensions have received disproportionate policy attention with little substantiated evidence, while other dimensions (such as CSAM and NCII) appear significant. Aside from CSAM and NCII, concern about the potential misuse of open foundation models is warranted, but substantive requirements for foundation model developers should be based on substantial empirical evidence.

5. Downstream interventions may be better suited to mitigating the risks of open foundation models. Based on the available evidence, it appears that the most severe

[3] David Thiel et al., “Generative ML and CSAM: Implications and Mitigations,” *Stanford Digital Repository*, June 24, 2023, https://purl.stanford.edu/jv206yg3793.

[4] Internet Watch Foundation, “How AI Is Being Abused to Create Child Sexual Abuse Imagery,” October 2023, https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

and likely marginal risks of open foundation models may be NCII and CSAM.[5] There have already been a number of reports of the harm caused by open models targeting specific individuals, though there are notable cases of NCII tied to closed models as well. Still, the best choke points here may lie downstream of model weights. Hardening downstream attack surfaces, and improving tracing of downstream model usage are important interventions that can mitigate the risks from both open and closed foundation models.[6]

6. The federal government should increase funding for risk assessment of foundation models. Reducing the uncertainty around the marginal risk of open foundation models will require further research and additional market surveillance. Government funding agencies should fund research to better understand the risks of both open and closed foundation models. Policymakers should also consider other actions to increase evidence generation and information sharing, such as disclosure requirements with disclosures directed toward both the government and the public.

<u>The unique benefits of open foundation models</u>

7. Open foundation models have distinct benefits relative to closed models. Foundation models with widely available weights present unique benefits that cannot be provided by closed models. Limiting the distribution of model weights would diminish the significant benefits provided by open models related to innovation, competition, transparency, and scientific discovery.

8. Open foundation models significantly advance innovation in AI. Open foundation models allow application developers to more easily adapt or fine-tune models on large proprietary datasets without the data protection concerns that come with transferring data to third parties. Customization expands the variety of applications that a foundation model can be incorporated into, empowering entrepreneurs. Importantly, open foundation models also enable concrete innovations, such as allowing models to be used in different languages.[7] The innovations provided by open foundation models are rapidly adopted by leading companies and small developers alike.[8]

9. Open foundation models underpin research on risk mitigation. Open foundation models might allow us to mitigate the very same risks that their skeptics decry. Model weights are essential for several forms of scientific research across AI interpretability, security, and safety. Broad access to foundation models bolsters the reproducibility of

[5] Note that the evidence base in this area is changing rapidly as additional relevant studies are being released on a near weekly basis.

[6] David Thiel, "Identifying and Eliminating CSAM in Generative ML Training Data and Models," *Stanford Internet Observatory*, December 23, 2023, https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf.

[7] Kunat Pipatanakul et al., "Typhoon: Thai Large Language Models," *arXiv*, December 21, 2023, https://arxiv.org/abs/2312.13951.

[8] Colin Raffel, "Building Machine Learning Models Like Open Source Software," *Communications of the ACM*, February 1, 2023, https://cacm.acm.org/opinion/building-machine-learning-models-like-open-source-software/.

scientific research and allows for better testing of safety guardrails.[9] But access to model weights alone is insufficient for certain types of safety research. Access to other assets, including training data and model checkpoints, are also necessary for advancing several forms of research related to risk, including research to understand biases and toxicity in data and models.[10]

10. Open foundation models enhance transparency. The 2023 Foundation Model Transparency Index indicates that developers of major open foundation models tend to be more transparent than their closed counterparts.[11] It found that open foundation model developers were more transparent on nine out of 13 major dimensions of transparency, often by a wide margin. Widely available model weights enable external researchers, auditors, and journalists to investigate and scrutinize foundation models more deeply. Broader scrutiny, including by underrepresented groups, helps reveal concerns missed by developers.[12] The availability of model weights alone does not guarantee full transparency on the upstream resources used to build the foundation model (e.g., data sources, labor practices, energy expenditure) or transparency on the downstream impact of the foundation model (e.g., adverse events and affected users), but closed foundation model developers are currently quite opaque in these areas.

11. Open foundation models help reduce algorithmic monoculture. Open models may also help combat algorithmic monoculture—where many actors in the AI ecosystem rely on the exact same algorithm.[13] By design, foundation models contribute to the rise of algorithmic monoculture as one model provides the basis for many different downstream applications.[14] This arrangement of AI development can contribute to individual harms like homogeneous outcomes,[15] systemic risks like security

[9] Sayash Kapoor and Arvind Narayanan, "OpenAI's Policies Hinder Reproducible Research on Language Models," *AI Snake Oil*, March 22, 2023, https://www.aisnakeoil.com/p/openais-policies-hinder-reproducible; Joon Sung Park et al., "Social Simulacra: Creating Populated Prototypes for Social Computing Systems," *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (74), October 2022, https://dl.acm.org/doi/10.1145/3526113.3545616, 1-18.

[10] Angelina Wang and Olga Russakovsky., "Directional Bias Amplification," *Proceedings of the 38th International Conference on Machine Learning* 139, July 2021, https://proceedings.mlr.press/v139/wang21t.html, 10882-10893.

[11] Rishi Bommasani et al, "The Foundation Model Transparency Index," https://crfm.stanford.edu/fmti/.

[12] Inioluwa Deborah Raji and Joy Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, January 2019, https://dl.acm.org/doi/10.1145/3306618.3314244, 429-35.

[13] Jon Kleinberg and Manish Raghavan, "Algorithmic Monoculture and Social Welfare," *Proceedings of the National Academy of Sciences* 118(22), May 25, 2021, https://www.pnas.org/doi/abs/10.1073/pnas.2018340118; Rishi Bommasani et al., "Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?," *Advances in Neural Information Processing Systems* 35, December 2022, https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html, 3663-3678.

[14] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," https://crfm.stanford.edu/assets/report.pdf#ethics; Rishi Bommasani et al., "Ecosystem Graphs: The Social Footprint of Foundation Models," https://arxiv.org/abs/2303.15772.

[15] Bommasani et al., "Picking on the Same Person."

vulnerabilities,[16] and economic harms like market concentration.[17] Open foundation models allow downstream developers to customize models to a great extent, allowing for greater differentiation in downstream model behavior. However, since most downstream development relies on a few open foundation models, the diversification of model behavior is limited.

12. Open foundation models promote competition in some layers of the AI stack. Given the significant capital costs of developing foundation models, broad access to model weights and greater customizability can also reduce market concentration by enabling greater competition in downstream markets.[18] However, open foundation models are unlikely to reduce market concentration in the highly concentrated upstream markets of computing and specialized hardware providers.[19]

13. Policymakers should consider the differential impact of AI policy on open foundation model developers. The government should be careful not to impose greater burdens on developers of open models as compared to well-resourced closed model developers. For example, certain proposals that contemplate liability for harms arising from downstream use of foundation models could chill the open foundation model ecosystem by exposing open model developers to risk that they cannot easily control.[20] More generally, open models are key to a competitive and vibrant AI ecosystem: imposing disproportionate burdens on their development will likely result in an industrial policy that concentrates economic power in a few developers of closed models. The United States is a global leader in open foundation models, and these models are on track to become a cornerstone of the U.S. digital economy.[21] AI regulation should aim to broaden access to safe models and systems, not restrict access to open ones.

We thank the Commerce Department and the NTIA for the opportunity to share our views, which are based on our scientific research in these areas. Please email nlprishi@stanford.edu, sayashk@princeton.edu, and kklyman@stanford.edu with any comments or questions.[22]

[16] Nicholas Carlini et al., “Poisoning Web-Scale Training Datasets Is Practical,” *arXiv*, February 20, 2023, https://arxiv.org/abs/2302.10149.

[17] Jai Vipra and Anton Korinek, “Market Concentration Implications of Foundation Models,” *arXiv*, November 2, 2023, https://arxiv.org/abs/2311.01550.

[18] Competition & Markets Authority, “AI Foundation Models Initial Report,” September 18, 2023, https://assets.publishing.service.gov.uk/media/650449e86771b90014fdab4c/Full_Non-Confidential_Report_PDFA.pdf.

[19] David Gray Widder, Sarah West, and Meredith Whittaker, “Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI,” *Concentrated Power, and the Political Economy of Open AI*, August 17, 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807.

[20] Rishi Bommasani et al., “Considerations for Governing Open Foundation Models,” *Stanford Institute for Human-Centered AI*, December 2023, https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf.

[21] Manuel Hoffmann et al., “The Value of Open Source Software,” *Harvard Business School*, January 1, 2024, https://www.hbs.edu/ris/Publication%20Files/24-038_51f8444f-502c-4139-8bf2-56eb4b65c58a.pdf.

[22] The authors of this response are writing in their personal capacities: These recommendations do not necessarily reflect the perspective of any of the organizations with which they are affiliated.

Sincerely,

Alondra Nelson

Arvind Narayanan

Caroline Meinhardt

Daniel E. Ho

Daniel Zhang

Dawn Song

Inioluwa Deborah Raji

Kevin Klyman

Marietje Schaake

Mihir Kshirsagar

Percy Liang

Peter Henderson

Rishi Bommasani

Rohini Kosoglu

Rumman Chowdhury

Sayash Kapoor

Seth Lazar

Shayne Longpre

Stefano Maffulli

Stella Biderman

Victor Storchan