DOCKET #240216-0052
RIN #0660-XC06
National Telecommunications and Information Administration
Herbert C. Hoover Building, 1401 Constitution Ave. NW
Washington, DC 20230

March 27, 2024

Re: NTIA's Request for Comment regarding Dual-Use Foundation Artificial Intelligence Models with Widely Available Model Weights as per Section 4.6 of the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

The Center for Democracy & Technology (CDT) respectfully submits these comments in response to NTIA's Request for Comment regarding the risks and benefits of, and potential policy approaches to, so-called "dual-use" foundation models for which the model weights are widely available, or as referred to in the RFC, "open foundation models" (OFMs).[1] Through this proceeding, required by section 4.6 of the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (the AI EO),[2] CDT is grateful to be able to share its perspective on how NTIA should advise the President on whether and how to regulate such models.

CDT is a nonprofit 501(c)(3) organization that works to advance civil rights and civil liberties in the digital age. Among our priorities, CDT advocates for the responsible and equitable design, deployment and use of new technologies such as artificial intelligence, and promotes the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems.

These comments build on a recent joint letter to the Commerce Department from CDT and a wide range of expert civil society organizations and academic scholars,[3] highlighting how substantial benefits may be lost, critical safety issues may be left under-addressed, and

[1] National Telecommunications and Information Administration, "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights," Federal Register, February 26, 2024, https://www.federalregister.gov/documents/2024/02/26/2024-03763/dual-use-foundation-artificial-intelligence-models-with-widely-available-model-weights. [perma.cc/8XKS-WMRH].
[2] See President Biden, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, October 2023, at sec. 4.6, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. [perma.cc/DDD3-VKWZ].
[3] Letter from Accountable Tech et al. to Secretary Gina Raimondo, March 25, 2024, https://cdt.org/insights/cdt-joins-mozilla-civil-society-orgs-and-leading-academics-in-urging-us-secretary-of-commerce-to-protect-ai-openness/. [perma.cc/8MJJ-Z2P2].

democratic values may be undermined, if the creation and publication of open foundation models are broadly targeted for regulation based on speculative risks.

Part I looks at the benefits of open foundation models, both by analogy to the history of open source software and by looking at recent AI developments, and concludes that they are likely substantial.

Part II considers the risks of open foundation models compared to closed models and other technologies like the internet — i.e., the *marginal* risks — and finds a need for more rigorous research into particular risks and their solutions.

Part III addresses policy approaches to open foundation models, with a focus on how the government can support critical field- and norm-building activities to clarify best practices around these risks, and how the government may need to tailor its policy interventions around open foundation models to satisfy the constraints of the First Amendment.

I. BENEFITS OF OPEN FOUNDATION MODELS

As previously highlighted in our attached joint letter and as discussed at length below, there are a wide variety of likely benefits from current and future open foundation models (OFMs) such as BERT, CLIP, Whisper, BLOOM, Pythia, Llama-2, Falcon, Stable Diffusion, Mistral, OLMo, Aya, and Gemma, as opposed to systems offered via the web or an API and without publication of key components including model weights. This conclusion is based on analogy to benefits from the vast open source software (OSS) ecosystem that has grown over the past three decades, as well as on developments around foundation models in just the past year and a half since OpenAI launched ChatGTP.

The Benefits of Open Source Software

"Open source" typically refers to software that has been released under a license that allows unrestricted use and modification of the software, including integrating it into other software tools or custom configurations. By contrast, not all OFMs are released under strictly "open source" licenses (although many are). Rather, safety considerations around AI have led to an emerging practice of publishing OFMs under what are sometimes called Responsible AI Licenses (RAIL) with some use restrictions to attempt to prevent undesirable types of deployments.[4] For this reason and others we will discuss later, "open source" and "open AI" are not always coextensive terms.

However, both open source software and OFMs allow for broader and more customized deployment than closed models that require direct permission from, payment to, or reliance on a

[4] BigScience, "BigScience RAIL License v1.0," May 19, 2022, https://huggingface.co/spaces/bigscience/license [perma.cc/48WK-QD6M]; Meta AI; "Llama 2 Community License Agreement," July 18, 2023. https://ai.meta.com/llama/license [perma.cc/L6PA-UTX8].

central service provider; and they further allow more in-depth study of the released components in order to improve, modify, and build on them. Therefore it is sensible to briefly look at the massive benefits over the past decades from open source generally before turning to those of OFMs specifically.

Over those decades, open source software has become a foundational element of society's entire software ecosystem. For example, imagine you are looking at an interesting blog post on your smartphone or laptop and choose to message a link to a friend:

- The web site you are viewing, like the vast majority of the web, is likely hosted on open-source Apache servers running on the open source Linux operating system;
- There's a very good chance that the web site uses the open source content management system WordPress; 43% of web sites use it.
- If you are reading it on a cellphone, you are likely reading it through the open source Chrome browser, using the open source Android operating system that runs on 63% of the world's billions of phones. Or you may be reading it on a Chromebook laptop, which is also running an open source operating system, Chromium. Thanks to these open source assets, a wide range of hardware providers around the world like Motorola, Samsung and Nokia — not just Google — have been able to more effectively compete with closed operating system ecosystems like Apple and Microsoft's.
- If you use private chat via the Signal, Whatsapp, or Google Messages apps to send the link to your friend, you are in turn using the open source Signal protocol for encrypted messaging.

The existence of open source alternatives like those above (as well as trailblazers like the open source Mozilla Firefox web browser that long prevented nearly complete dominance of the PC browser market by Microsoft's Internet Explorer) has led to the existing level of competition in the internet software and computing hardware space.

Open source's impact extends essentially to all software: today, 96% of all code bases include open-source software,[5] and GitHub, the biggest platform for the open-source community, is used by more than 100 million developers worldwide.[6] Translating that into economic impact, a recent analysis from Harvard Business School found that without open source, firms would likely have to spend 3.5 times more on software,[7] while in a survey businesses themselves similarly estimated they would have to spend 4 times more.[8] Executives of Fortune 500 companies

[5] Synopsis. "Open Source Security and Analysis Report (OSSRA)," 2024, https://www.synopsys.com/software-integrity/engage/ossra/ossra-report. [perma.cc/F7QQ-UL3D].
[6] Thomas Dohmke, "100 Million Developers and Counting - the GitHub Blog," The GitHub Blog, January 25, 2023, https://github.blog/2023-01-25-100-million-developers-and-counting. [perma.cc/9KXH-QE4A].
[7] Manuel Hoffmann, Frank Nagle, and Yijian Zhou, "The Value of Open Source Software," *Social Science Research Network*, January 1, 2024, https://doi.org/10.2139/ssrn.4693148. [perma.cc/73S7-L6NX].
[8] Henry Chesbrough, "Measuring the Economic Value of Open Source," The Linux Foundation, March 2023 March, 2023, https://www.linuxfoundation.org/research/measuring-economic-value-of-os. [perma.cc/K6TY-5VKD].

highlight the cost savings, faster development, and better interoperability offered by open source software.[9]

Open source has proven economically valuable not only to users but to developers. For example, the stock market responds positively when technology firms release new open source technologies,[10] while startups with higher levels of contribution to GitHub tend to see increases in their funding and valuations.[11] And the European Commission estimated that approximately a €1 billion investment in open source software by European companies in 2018 resulted in an impact on the European economy of between *€65 and €95 billion*. Considering those kinds of numbers and the safety benefits of open source that we'll discuss later, it is unsurprising that a recent bipartisan legislative proposal recognized that "a secure, healthy, vibrant, and resilient open source software ecosystem is crucial for ensuring the national security and economic vitality of the United States."[12]

Open source is admittedly not a silver bullet that will solve all competition problems, and it can in fact be leveraged by larger players as a means of solidifying influence on the technical ecosystem. For example, Google's forays into open source mobile operating systems and browsers have enhanced that company's structural power, while the main competitors it was challenging remain among the most capitalized companies in the world. However, it is not hard to imagine how much more concentrated and immovable the incumbent positions of Apple and Microsoft would be now if they had not faced open source-driven challenges from Google and the many hardware vendors they enable.

The past of open source is not a promise that open AI generally or open foundation models in particular will have all of the same kinds of positive effects, however. So we must also examine the benefits of OFMs in particular. We will highlight three major categories of likely benefits: distributing power, catalyzing innovation, and ensuring transparency.[13]

---

[9] Id.

[10] Wei Yang, "How Can Open Source Technology Ecosystem Create Value? Evidence From Investors' Reactions to Firms' GitHub Code Releases," *Social Science Research Network*, April 30, 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4433433. [perma.cc/G5BH-NNT6].

[11] Nataliya Langburd Wright, Frank Nagle, and Shane Greenstein, "Contributing to Growth? The Role of Open Source Software for Global Startups," vol. 52, January 1, 2024, https://www.hbs.edu/ris/Publication%20Files/24-040_69bae20b-2026-4089-b76c-07b8a8cc48d4.pdf. [perma.cc/E5AE-222C].

[12] Gary C. Peters, "S.4913 - Securing Open Source Software Act of 2022," September 28, 2022, https://www.congress.gov/bill/117th-congress/senate-bill/4913 [perma.cc/RXX8-FMRY].

[13] These three categories are taken from Rishi Bommasani et al., "Considerations for Governing Open Foundation Models," *Stanford Institute for Human-Centered Artificial Intelligence*, December 13, 2023, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models [perma.cc/L93F-BCDM]. Several of the same authors further break down benefits into a more detailed set of five categories in a later paper, see Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models," *arXiv*, February 27, 2024, https://arxiv.org/abs/2403.07918 [perma.cc/P9N9-9QSQ].

Distributing Power (Both in the Market and the Culture)

The economic benefits of generative AI driven by foundation models are expected to be enormous, with Bloomberg predicting a market of $1.3 trillion by 2032.[14] However, recent research indicates that the market structure for closed foundation models has a tendency toward concentration, including vertical integration, in large part due to the high costs of compute infrastructure for training.[15]

In addition to massive concentration of economic power, a foundation model market dominated by a handful of closed systems carries other risks. For example, when many different decisionmakers and service providers rely on the same systems, there can be a trend toward "algorithmic monoculture" whereby systemic exclusion of individuals or groups in AI-driven decisionmaking occurs across the ecosystem.[16] There is also the risk of actual monoculture, where a handful of companies decide what knowledge and expression is allowed through this powerful new layer of information technology, raising the specter of undue power over politics and culture. This is a fraught issue, still under debate in the context of social media companies; and that debate — which should be of concern regardless of one's politics, left or right — is now extending to the acceptable use and content moderation efforts of closed foundation models.[17] In the social network realm, we've seen the beginning of a move toward decentralized social networks built on open standards to allow for a range of different types of social networks serving different needs, communities, and social norms;[18] open foundation models similarly provide a decentralized alternative to the concentration of power and decisionmaking in a handful of closed providers.

---

[14]Oktavia Catsaros, "Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds," *Bloomberg Intelligence*, June 1, 2023, https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/. [perma.cc/3Z42-6J2B].

[15] Jai Vipra and Anton Korinek, "Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT," *Center on Regulation and Markets at Brookings*, September 7, 2023, https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt [perma.cc/D5TP-KCKV]; David Gray Widder, Sarah Myers West, and Meredith Whittaker, "Open (for Business): Big Tech, Concentrated Power, and the Political Economy of Open AI," *Social Science Research Network*, January 1, 2023, https://doi.org/10.2139/ssrn.4543807 [perma.cc/5ERM-3E39].

[16] Rishi Bommasani et al., "Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?," *arXiv*, November 25, 2022, https://arxiv.org/abs/2211.13972. [perma.cc/F7JB-3AK3]

[17] Nitasha Tiku, Kevin Schaul, and Szu Yu Chen, "AI generated images are biased, showing the world through stereotypes," *Washington Post*, November 1, 2023, https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes [perma.cc/3NG5-NDLY]; Dan Milmo and Alex Hern, "'We definitely messed up': why did Google AI tool make offensive historical images?," *The Guardian*, March 8, 2024, https://www.theguardian.com/technology/2024/mar/08/we-definitely-messed-up-why-did-google-ai-tool-make-offensive-historical-images [perma.cc/27S9-4AR6]; Fábio Yoshio Suguri Motoki, Valdemar Pinho Neto, and Víctor Rodrigues, "More human than human: measuring ChatGPT political bias," *Public Choice* 198, no. 1-2 (August 17, 2023): 3-23, https://doi.org/10.1007/s11127-023-01097-2 [perma.cc/27S9-4AR6].

[18] Roel Roscam Abbing, Cade Diehm, and Shahed Warreth, "Decentralised Social Media," *Internet Policy* Review, February 20, 2023, https://policyreview.info/glossary/decentralised-social-media [perma.cc/TC4M-SR8F].

Based on the history of open source software, one would also expect OFMs to enable faster, cheaper diffusion of foundation model technology to startups and other businesses large and small, as well as other developer and user communities around the world. And, so far, that is exactly what is occurring. As the UK's Competition and Markets Authority has noted, "[a]t present a mix of open and closed-source foundation models are available and competing. This is allowing a range of firms to invest in and develop foundation models and as a result we are already seeing deployment of these foundation models in a growing range of applications across the economy."[19] For example, as of September 23, 2023 — half a year ago now — Meta reported that already tens of thousands of entrepreneurs and startups were using Llama-2,[20] while as of December 7, 2023, Meta reported that their Llama models had been downloaded over 100 million times.[21] Google's new suite of Gemma models (7B, 7B-IT, 2B, & 2B-IT), meanwhile, were downloaded over one million times in the last month from the Hugging Face platform.[22]

The demand for OFMs is being seen in a wide range of contexts. Large companies such as Dell and Wells Fargo are starting to use them to help with internal knowledge management and internal software coding, with Dell's SVP for AI Strategy noting: "A lot of customer[s] are asking themselves: "Wait a second, why am I paying for [a] super large model that knows very little about my business? Couldn't I just use one of these open-source models, and by the way, maybe use a much smaller, open-source model for that (information retrieval) workflow?"[23] OFMs are now being offered and widely used on the cloud platforms of Microsoft, AWS and Google, and consultants like McKinsey are using them to build applications for their clients.[24] Small technology firms and startups with fewer resources are depending heavily on the availability of free pre-trained models that they can adapt to their applications,[25] and limitations on the availability of such models could disproportionately impact those small competitors.

CDT's own research in interviews with deployers who are leveraging foundation models offers similar conclusions: they stress that frequent changes in closed model APIs, model versions, or terms of service make navigating contracts with clients and maintaining stable builds more difficult.[26] They also note that safety guardrails in the foundation models can make stress testing their own applications and creating robust, application-specific safety checks challenging.

---

[19] Competitions and Markets Authority. "AI Foundation Models: Initial Review," *GOV.UK*, February 28, 2024, https://www.gov.uk/cma-cases/ai-foundation-models-initial-review. [perma.cc/V6H3-3KLD].
[20] Joe Spisak and Sergey Edunov, "The Llama Ecosystem: Past, Present, and Future," *Meta AI* (blog), September 27, 2023, https://ai.meta.com/blog/llama-2-updates-connect-2023 [perma.cc/USS2-5JWY].
[21] Meta, "Introducing Purple Llama for Safe and Responsible AI Development," *Meta Newsroom*, December 12, 2023, https://about.fb.com/news/2023/12/purple-llama-safe-responsible-ai-development. [perma.cc/4M6U-P9VG].
[22] "Google," HuggingFace, https://huggingface.co/google. [perma.cc/4G63-894P].
[23] Matt Marshall, "How Enterprises Are Using Open Source LLMs: 16 Examples," *VentureBeat*, February 2, 2024, https://venturebeat.com/ai/how-enterprises-are-using-open-source-llms-16-examples. [perma.cc/PVL3-X6LX].
[24] Id.
[25] Amy A. Winecoff and Elizabeth Anne Watkins, *Artificial Concepts of Artificial Intelligence*, *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, https://doi.org/10.1145/3514094.3534138. [perma.cc/BQ5U-N9JV].
[26] On file with the author, research publication forthcoming.

Relying on closed models is especially fraught when efforts over API through a closed foundation model provider to fine-tune for a particular application are not portable to another foundation model provider or cloud host, thereby further reducing competition and increasing vendor lock-in. In contrast, an open model can be hosted purely internally, or externally with a wide variety of hosts, and can be moved between them.

The customization enabled by OFMs also can speed dispersion of the technology in culturally relevant forms, throughout the globe including the global south. Right now, most major commercial OFMs are English-dominated, but LLaMA-2 has enabled researchers to train models for low-resource languages such as those spoken in Southeast Asia, in ways that better reflect local cultural norms, values, and legal considerations.[27] The availability of open, multilingual language models such as the open access Aya LLM[28] can similarly broaden access to LLMs globally, and promote further innovation in geographic areas that are typically underserved by the dominant closed systems.

In sum, the course of OFM development over just the course of the past couple of years demonstrates that it is already a powerful competitive alternative to closed foundation models, and likely will continue to be unless artificially stymied by restrictions on publication.

Catalyzing Innovation (Both in AI and Other Fields)

As described above, OFMs are already driving innovation across the ecosystem as tens or hundreds of thousands of businesses begin adapting model capabilities to their own use cases and customer needs in a wide variety of contexts. And the spread of these technologies will also help science advance in a wide range of fields. But open source is also at the root of much AI innovation today.

Open source machine learning software including OFMs can and already has driven significant advances in AI technology. Indeed, the current flourishing of generative AI and foundation model technology would not have been possible without open research. For example, the 2017 paper that originated the technology that underlies today's LLMs, transformer networks, was open research with open code and data,[29] as was the research paper that debuted one of the most popular early language models,[30] work that enabled the current closed foundation models to design their systems. Without this open research — including the release of what at the time were the most sophisticated or "frontier" language models — and open source ML development

[27] Xuan-Phi Nguyen et al., "SeaLLMs -- Large Language Models for Southeast Asia," *arXiv*, December 1, 2023, https://arxiv.org/abs/2312.00738 [perma.cc/5UW5-MZLX]; Jun Zhao et al., "Llama Beyond English: An Empirical Study on Language Capability Transfer," *arXiv*, January 2, 2024, https://arxiv.org/abs/2401.01055 [perma.cc/YPR5-7JX5].
[28] Ahmet Üstün et al., "Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model," *arXiv*, February 12, 2024, https://arxiv.org/abs/2402.07827. [perma.cc/LF7V-Z2CJ].
[29] Ashish Vaswani et al., *Attention Is All You Need*, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, vol. 30, 2017, https://arxiv.org/abs/1706.03762. [perma.cc/TMN4-GX84].
[30] Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, October 11, 2018, https://arxiv.org/abs/1810.04805. [perma.cc/W25M-DVNY].

frameworks like Pytorch and TensorFlow, today's closed models would not even exist. In fact, it is fair to conclude that almost *all* ML development has heavily relied on advances in open tools and open models.[31] Meanwhile, open research is also facilitating new progress in finding ways to enhance the safety of models, including research to help ensure interventions remain durable or can be enhanced even after models are released.[32]

Open models have not only been critical catalysts for the development of foundation models, both open and closed. They have also been a key ingredient for creating smaller, more efficient, and customized models with little cost that can rival larger foundation models. For example,

- Vicuna-13B is an open-source chatbot that was developed by fine-tuning Llama based on user-shared conversations with ChatGPT. At the time of release, the model showed strong performance in preliminary assessments compared to ChatGPT and Bard despite costing only around $300 to train.[33]

- Researchers used the model weights of Mistral 7B, a 7.3 billion parameter model released under an open source license by the startup Mistral,[34] to decrease the computational power required for fine-tuning the model for downstream tasks by a factor of ten.[35]

- Alpaca 7B, a language model developed by fine-tuning Llama for instruction following, demonstrated qualitatively similar performance to GPT-3.5 while costing under $600 to train.[36]

- Koala-13B, a language model developed by fine-tuning Llama on dialogue data scraped from the web, demonstrated better performance than Alpaca and similar performance to ChatGPT in preliminary assessments. Koala cost around $100 to train.[37]

---

[31] Max Langenkamp and Daniel N. Yue, *How Open Source Machine Learning Software Shapes AI*, *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, 2022, https://doi.org/10.1145/3514094.3534167. [perma.cc/PX3S-DUGM].
[32] Eric Mitchell et al., "Fast Model Editing at Scale," *arXiv*, October 21, 2021, https://arxiv.org/abs/2110.11309 [perma.cc/X29Y-WFU2]; Kevin Meng et al., *Locating and Editing Factual Associations in GPT*, *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, vol. 35, 2022, https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html [perma.cc/PCT3-EMUP].
[33] The Vicuna Team. "Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90%* ChatGPT Quality." *LMSYS Org (blog)*, March 30, 2023. https://lmsys.org/blog/2023-03-30-vicuna. [perma.cc/BG2M-TLGE]
[34] Albert Q. Jiang et al., "Mistral 7B," *arXiv*, October 10, 2023, https://arxiv.org/abs/2310.06825. [perma.cc/TZT5-UXXY].
[35] James Liu et al., "BitDelta: Your Fine-Tune May Only Be Worth One Bit," *arXiv*, February 15, 2024, https://arxiv.org/abs/2402.10193. [perma.cc/4QWF-73W7].
[36] Rohan Taori et al., "Alpaca: A Strong, Replicable Instruction-Following Model," *Stanford Center for Research on Foundation Models*, March 13, 2023, https://crfm.stanford.edu/2023/03/13/alpaca.html. [perma.cc/V5GY-4GBQ].
[37] Ritwik Gupta et al., "Koala: A Dialogue Model for Academic Research," *The Berkeley Artificial Intelligence Research Blog*, April 3, 2023, https://bair.berkeley.edu/blog/2023/04/03/koala. [perma.cc/C956-J3ZN].

Importantly, the innovation in developing smaller and more powerful models, often based directly on much larger models, is not just important in terms of competition and innovation. It is also important because some models such as Mistral 7B are now small enough to run locally on an end-user's laptop or even a phone, mitigating the need for a cloud-based provider at all.[38] This brings a number of benefits to consumers and society, including privacy (data need not travel the internet or go to anyone's cloud), greater speed, and no need for access to the internet or hosting at a data center (particularly relevant given the environmental impacts of the AI-driven demand for data centers[39]).

Furthermore, OFMs enable a variety of AI research not enabled by closed foundation models,[40] including research around AI interpretability methods,[41] security, model training and inference efficiency,[42] and the public development of robust watermarking techniques.[43]

Finally, faster dispersion of open models means faster advancement of scientific research across fields, and doing that research with open models can help address issues of scientific reproducibility and verifiability. For example, in a meta-analysis of over 400 papers addressing the utility of AI in imaging for COVID-19 patient care, the highest quality papers almost all relied on open pretrained models, suggesting that the availability of open source models may be crucial to future AI-enabled medical advancements.[44] As the researchers concluded, "[g]iven the global, unprecedented public health challenge caused by COVID-19, we strongly encourage medical researchers to follow the trends toward open-source development in the field of ML."[45] This admonition could just as well apply to other urgent public needs and scientific research in general, but the need for such research and deployment transparency is most especially important in sensitive use cases including medicine, as we'll explore more below.

---

[38] Jennie Rose, "How to Run Llama 2 Locally: The Ultimate Guide for Mac, Windows, and Mobile Devices," *Cheatsheet.Md*, March 17, 2024, https://cheatsheet.md/llm-leaderboard/how-to-install-local-llama [perma.cc/ZY5F-9WC8]; Chris McKay, "How to Get Started With Mistral 7B," *Maginative*, September 29, 2023, https://www.maginative.com/article/how-to-get-started-with-mistral [perma.cc/ZZN7-FWSW].

[39] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat, "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model," *Journal of Machine Learning Research 2* 24 (2023), https://www.jmlr.org/papers/v24/23-0069.html [perma.cc/64UB-8LP9]; David Patterson et al., "Carbon Emissions and Large Neural Network Training," *arXiv*, April 21, 2021, https://arxiv.org/abs/2104.10350 [perma.cc/5WN2-J8XS].

[40] Bommasani et al., 2023, *supra* note 13.

[41] Kevin Clark et al., "What Does BERT Look at? An Analysis of BERT's Attention," *arXiv*, June 11, 2019, https://arxiv.org/abs/1906.04341. [perma.cc/PH9S-BB4S].

[42] Tim Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv.org, May 23, 2023, https://arxiv.org/abs/2305.14314 [perma.cc/66JX-ALXM]; Sunny S. Sanyal et al., "Early Weight Averaging Meets High Learning Rates for LLM Pre-training," *arXiv*, June 5, 2023, https://arxiv.org/abs/2306.03241 [perma.cc/QEW7-M86W].

[43] John Kirchenbauer et al., "A Watermark for Large Language Models," *arXiv.Org*, January 24, 2023, https://arxiv.org/abs/2301.10226. [perma.cc/7ZRX-7876].

[44] Jannis Born et al., "On The Role of Artificial Intelligence in Medical Imaging of COVID-19," *Patterns* 2, no. 6 (April 30, 2021): 100269, https://doi.org/10.1016/j.patter.2021.100269. [perma.cc/5TK8-FMCR].

[45] Id.

Ensuring Transparency (and Security and Accountability)

"With enough eyeballs," an old saying in open source software development goes, "all bugs are shallow."[46] By opening itself to scrutiny from a global community of professional and amateur developers, opening software has typically been viewed as a benefit rather than a detriment. Through open sourcing, one can essentially recruit an army of white hat hackers to help counter the army of black hat hackers that will be trying to break the thing.[47] This is considered a helpful strategy because a strategy of secrecy — or "security by obscurity" — is less effective when a target can be repeatedly and endlessly probed by hostile actors (e.g. is connected to the internet).[48] Openness enables, for example, powerful pro-security efforts like the OSS-Fuzz project, which continuously scans across hundreds of open-source projects for vulnerabilities.[49]

It is because of these security benefits, amongst others such as cost and customizability, that the federal government strongly prefers open source software, even as it works to further improve its security. For example, the Digital Services Playbook urges government offices to "default to open,"[50] NIST has long recommended for secure systems a principle of "open design", i.e., that "security should not depend on the secrecy of the implementation or its components,"[51] and the Department of Defense site on open source highlights that "[c]ontinuous and broad peer review, enabled by publicly available source code, improves software reliability and security through the identification and elimination of defects that might otherwise go unrecognized by the core development team."[52]

Of course, AI systems are not exactly the same as other software systems: although various components of a foundation model are software, the source code of which can be examined by programmers, the weights themselves — essentially a massive multidimensional database of relationships between the tokens the model was trained on — are not directly human-readable. Furthermore, the "safety" issues that many may want to test a foundation modell for are not typically traditional security vulnerabilities in code but rather poor or harmful model behavior or generated outputs. Finally, to the extent such issues are discovered in an OFM, patches via

---

[46] Eric S. Raymond, "The Cathedral and the Bazaar," *Knowledge, Technology & Policy* 12, no. 3 (September 1, 1999): 23-49, https://doi.org/10.1007/s12130-999-1026-0. [perma.cc/7Z3G-V9CB]
[47] Steven Weber, "The Success of Open Source," in *Harvard University Press eBooks*, 2004, https://doi.org/10.4159/9780674044999. [perma.cc/GMD2-V5LU]
[48] Peter P. Swire, "A Model for When Disclosure Helps Security: What Is Different About Computer and Network Security?," *Journal on Telecommunications and High Technology Law* 163 (2004), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=531782. [perma.cc/5B9R-XYKJ]
[49] Dongge Liu et al., "AI-Powered Fuzzing: Breaking the Bug Hunting Barrier," Google Online Security Blog, August 16, 2023, https://security.googleblog.com/2023/08/ai-powered-fuzzing-breaking-bug-hunting.html. [perma.cc/B3T6-TPNA]
[50] U.S. Digital Service. "The Digital Services Playbook," n.d. https://playbook.cio.gov/. [perma.cc/2VYD-GM9G]
[51] Karen Scarfone et al., NIST Special Publication 800-123, *Guide to General Server Security* (2008), https://csrc.nist.gov/pubs/sp/800/123/final, at 4. [perma.cc/JCY7-3ANY]
[52] U.S. Department of Defense. "Open Source Software FAQ." U.S. Department of Defense Chief Information Officer, October 28, 2021, https://dodcio.defense.gov/Open-Source-Software-FAQ. [perma.cc/KW2U-U739]

fine-tuning or other interventions may not be universally adopted, particularly by malicious actors. An unfixed model version could still be available to those who seek to use it — and because patches are not just security fixes but suppression of concepts or capabilities that malicious actors may want to deploy offensively, the malicious actors may be motivated to continue using unmitigated or maliciously modified versions of the model that facilitate their offensive goals. Therefore not all of the assumptions around security and open source may hold in the case of AI, and more study in this area is warranted.[53]

That said, we are already seeing examples of openness — both in the data on which OFMs are trained, and in the models themselves — contributing to security, safety, and other critical research beneficial to both open *and* closed systems, as well as instances where similar research into closed systems has been stymied. For example, researchers' discovery of child sexual abuse material (CSAM) in the LAION data set that is regularly used by both open and closed models would not have been possible if that data set were not open.[54] Similarly, foundational research on the fragility of fine-tuned guardrails in both open and closed models — not only when subject to deliberate attack but even when subject to benign fine-tuning for other purposes — was based on use and examination of OFMs and would not have been possible otherwise.[55] Meanwhile, recent state-of-the-art work in auditing of closed models also leverages OFMs: for example, the Llama-based Vicuna model trained on GPT-4 outputs has enabled researchers to identify attack vectors that can then be tested on GPT-4 itself.[56]

By contrast, as already mentioned, a number of types of general AI research and auditing cannot be fully conducted with closed foundation models.[57] A recent and relevant example of the importance of a broader community testing for AI harms is new research on bias in LLMs, which revealed that even when models do not exhibit overt racial bias in their responses to users, they can contain and exhibit more subtle biases in consequential domains such as employment and criminal justice when prompts contain African American English (AAE) as opposed to Standard American English (SAE). For example, GPT-4 was shown to be more likely to suggest that defendants be sentenced to death when they provide statements in AAE. Even more troubling, this covert bias was seen to *increase* rather than decrease with the size of models, and while it is possible for human feedback training to mitigate this covert bias based on dialect, it also can exacerbate it by teaching models to superficially conceal overt racial biases while still containing covert ones.[58]

---

[53] See Toby Shevlane and Allan Dafoe, "The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?," *arXiv*, December 27, 2019, https://arxiv.org/abs/2001.00463. [perma.cc/387A-MMA4]

[54] David Thiel, "Identifying and Eliminating CSAM in Generative ML Training Data and Models" (Stanford Internet Observatory Cyber Policy Center, December 20, 2023), https://purl.stanford.edu/kh752sm9123. [perma.cc/UFU8-HRAR]

[55] Xiangyu Qi et al., "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!," *arXiv*, October 5, 2023, https://arxiv.org/abs/2310.03693. [perma.cc/G6R8-XHQT]

[56] Andy Zou et al., "Universal and Transferable Adversarial Attacks on Aligned Language Models," *arXiv*, July 27, 2023, https://arxiv.org/abs/2307.15043. [perma.cc/ED4C-Q9ZE]

[57] *Supra* at notes 40-43.

[58] Valentin Hofmann et al., "Dialect Prejudice Predicts AI Decisions About People's Character, Employability, and Criminality," *arXiv*, March 1, 2024, https://arxiv.org/abs/2403.00742

This research is notable not only for the severity of the problem it highlights, nor for its stark demonstration of why these systems are not ready to make important decisions about human lives. It demonstrates the inherent limitation in testing closed models for harms: the researchers were not able to use their entire battery of tests and fully complete their research on GPT-4 compared to other language models due to the closed nature of GPT-4.[59]

The inability to effectively audit closed foundation models compared to OFMs is a systemic problem.[60] As recent research explains, audits can be broken into black-box methods (auditors can only access inputs and outputs, not internal model weights) and white-box methods (auditors get unrestricted access to internal workings, including model weights). Closed foundation models can only be audited using black-box methods, but OFMs can also be audited with white-box methods. Additional contextual materials about a model's development, whether closed or open, can also enable additional "outside the box" auditing.

Black-box audits assess model characteristics using test sets as inputs or otherwise trying to find inputs that lead to harmful outputs.[61] However, these approaches are fundamentally limited in how they are able to identify harms, since these methods amount to searching for problems only in an exploratory manner or where robust and reliable evaluations exist. Black-box audit methods also fail to allow for auditors to gain a generalized understanding of how a system works and what its shortcomings might be. Many larger models are actually combinations of a variety of different expert models; but without the ability to understand or access those components separately, a black-box study could easily and inadvertently focus on one part of a model while unknowingly overlooking other parts. Furthermore, because black-box audits are necessarily based entirely on responses to inputs, minor differences in the content of those inputs can lead to widely different results, making those results less reliable and harder to

[perma.cc/C88W-YAUG]. For a summary, see also Elizabeth Gibney, "Chatbot AI makes racist judgements on the basis of dialect," *Nature* 627, no. 8004 (March 13, 2024): 476-77, https://doi.org/10.1038/d41586-024-00779-1 [perma.cc/23P2-D37E].

[59] Id. (Specifically, the researchers compared probabilities of adjectives related to African Americans, but could not conduct this analysis since it requires access to the probabilities for all adjectives, which GPT-4's API only provides for the top five predicted tokens; researchers could also not compute model perplexity using the OpenAI API so excluded GPT-4 from analyses based on perplexity).

[60] Stephen Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," *arXiv*, January 25, 2024, https://arxiv.org/abs/2401.14446 [perma.cc/Q29G-XVB2]; see also Victor Ojewale et al., "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling," arXiv.org, February 27, 2024, https://arxiv.org/abs/2402.17861 [perma.cc/LML4-K465].

[61] E.g., Alexander Wei, Nika Haghtalab, and Jacob Steinhardt, "Jailbroken: How Does LLM Safety Training Fail?," *arXiv*, July 5, 2023, https://arxiv.org/abs/2307.02483 [perma.cc/4WW9-DRXF]; Hofmann et al., 2024, *supra* note 58; Jesutofunmi A. Omiye et al., "Large Language Models Propagate Race-based Medicine," *Npj Digital Medicine* 6, no. 1 (October 20, 2023), https://doi.org/10.1038/s41746-023-00939-z [perma.cc/L9KW-8JKQ].

reproduce.[62] Unreliable measurements in turn makes it that much harder for auditors to assess the source of problems[63] and to recommend specific mitigations to improve the model.[64]

White-box auditing methods, in contrast, allow for testing to be more directly guided by auditors and more efficient in locating problems than the unguided, trial-and-error probing involved in black-box methods. In particular, methods unique to white-box models such as gradient-based optimization allow for finding attack vectors in vision systems,[65] and to a lesser extent language ones.[66] Gradient-based techniques (again, unique to white-box models) also could allow auditors to better understand a model's individual decisions by highlighting what part of a given input (e.g. a prompt, or an image to classify) is most relevant to the generation of a given output.[67] White-box access also better enables a range of novel auditing methods including "methods based on local search, rejection sampling at scale, Langevin dynamics, evolutionary algorithms, and reinforcement learning."[68] This range of white-box methods allow auditors to probe more effectively for new capabilities and to test for jailbreaks.[69]

Furthermore, in contrast to white-box audits, closed models that are only open to black-box methods can control who is allowed to probe or audit their systems and how. Over 300 researchers have complained in an open letter that "AI companies have already suspended researcher accounts and even changed their terms of service to deter some types of evaluation."[70] This creates a chilling effect on independent research of closed models, including

---

[62] Moran Mizrahi et al., "State of What Art? A Call for Multi-Prompt LLM Evaluation," *arXiv*, December 31, 2023, https://doi.org/10.48550/arxiv.2401.00595 [perma.cc/5JLL-8EFV]; Norah Alzahrani et al., "When Benchmarks Are Targets: Revealing the Sensitivity of Large Language Model Leaderboards," *arXiv*, February 1, 2024, https://arxiv.org/abs/2402.01781 [perma.cc/XK73-QNGZ].

[63] Amy Winecoff and Miranda Bogen, "Trustworthy AI Needs Trustworthy Measurements - Center for Democracy and Technology," Center for Democracy and Technology, March 6, 2024, https://cdt.org/insights/trustworthy-ai-needs-trustworthy-measurements. [perma.cc/NX6E-WXXM]

[64] Song Wang et al., "Knowledge Editing for Large Language Models: A Survey," *arXiv*, October 24, 2023, https://arxiv.org/abs/2310.16218. [perma.cc/74EB-M8D4]

[65] Ian J. Goodfellow,, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv*, December 20, 2014, https://arxiv.org/abs/1412.6572. [perma.cc/6LPX-CNK3]

[66] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh, "Universal Adversarial Triggers for Attacking and Analyzing NLP," *arXiv*, August 20, 2019, https://arxiv.org/abs/1908.07125. [perma.cc/MV7A-BTVW]

[67] Arun Das and Paul Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," *arXiv*, June 16, 2020, https://arxiv.org/abs/2006.11371. [perma.cc/2BL2-9NMT]

[68] Stephen Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," *arXiv*, January 25, 2024, https://arxiv.org/abs/2401.14446. [perma.cc/FU5Z-9URX]

[69] e.g., Alain Guillaume and Yoshua Bengio, "Understanding Intermediate Layers Using Linear Classifier Probes," *arXiv*, October 5, 2016, https://arxiv.org/abs/1610.01644 [perma.cc/3QFV-FNWA]; Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, and Levent Sagun, "Fairness Indicators for Systematic Assessments of Visual Feature Extractors," *arXiv*, February 15, 2022. https://arxiv.org/abs/2202.07603 [https://perma.cc/K87Y-9E9Z].

[70] Letter from Arvind Narayanan et al., "A Safe Harbor for Independent AI Evaluation," March 2024, https://sites.mit.edu/ai-safe-harbor/. [perma.cc/WC5C-ZYMN]

research to detect potential harms that developers may have overlooked, due to fear of loss of access to the model or even legal reprisals.[71]

The need for white-box and outside-the-box auditing beyond closed box methods is all the more evident when considering that foundation models may be applied in contexts where consequential decisions about people are being made, which puts them at risk of being subjected to systemic biases. We've already described the covert racial bias study which has troubling implications for a wide variety of deployment contexts. Meanwhile, experts and practitioners in other sensitive fields like medicine and law are also calling for more reliance on open models to ensure greater control and better decisionmaking,[72] especially in the face of studies (e.g.) demonstrating racial bias in medical decisionmaking by closed models.[73]

Finally, the availability of OFMs will necessarily assist in the education and training of new computer scientists in the particulars of how to develop, test, and deploy foundation models, opportunities that will be much more limited if students can only engage in white-box testing of less advanced open models (where learnings may not be transferable to larger or more advanced systems) and black-box testing with a few closed foundation model systems. The usefulness for researchers of the emerging National AI Research Resource will also turn on such access; the alternative is a less useful program that relies on closed services donating access to their models, and/or the government paying them. We should not build an open national research resource in a manner that drives researchers and taxpayer dollars to rely on closed services, rather than openly available and testable assets like OFMs.

II. MARGINAL RISKS OF OPEN FOUNDATION MODELS

In evaluating the risks of OFMs, we must consider them in comparison to the existing risks enabled by closed models, by access to existing technologies such as the internet, and by smaller models that carry similar risks but for which controlling proliferation would be much harder if not impossible. In other words, we must consider the *marginal* risk of OFMs.[74]

---

[71] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, et al., "A Safe Harbor for AI Evaluation and Red Teaming," *arXiv*, March 7, 2024, https://arxiv.org/abs/2403.04893. [perma.cc/QS2B-AZTC]

[72] Augustin Toma, Senthujan Senkaiahliyan, Patrick R. Lawler, Barry B. Rubin, and Bo Wang, "Generative AI Could Revolutionize Health Care — but Not if Control Is Ceded to Big Tech," *Nature* 624, no. 7990 (November 30, 2023): 36-38, https://doi.org/10.1038/d41586-023-03803-y [perma.cc/2EN4-ESJF]; Anthony Dang, "The Open Advantage: Winning the Adversarial Battle With Open-Source Models," *Social Science Research Network*, January 1, 2023, https://doi.org/10.2139/ssrn.4651571 [perma.cc/M9KX-KZE8].

[73] Jesutofunmi A. Omiye, Jonathan Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou, "Large Language Models Propagate Race-based Medicine," *Npj Digital Medicine* 6, no. 1 (October 20, 2023), https://doi.org/10.1038/s41746-023-00939-z. [perma.cc/8RMQ-FVMV]

[74] Kapoor et al., 2024 *supra* note 13. According to the framework in this paper, to effectively gauge marginal risk, one must not only identify the risk, but also existing risks absent OFMs, existing defenses absent OFMs, actual evidence of marginal risk, new defenses that could be used, and the assumptions underlying the analysis. Existing literature on OFM risks are severely lacking in several of these categories, therefore "[a]cross several misuse vectors (e.g. cyberattacks, bioweapons), we find that

It is critical to marshall meaningful evidence demonstrating the likelihood and severity of actual marginal risks before policymaking, to ensure that proposed solutions are a good fit for the problem. And while there are a number of organizations vigorously advocating the seriousness of certain risks, as the bipartisan leadership of the House Committee on Science, Space, and Technology observed, research findings from this community are "often self-referential and lack the quality that comes from revision in response to critiques by subject matter experts."[75]

This is not to suggest that these expert advocates are not sincere in their concerns or that their work doesn't serve as an important foundation upon which more research can be based; we believe that the risks they raise must be considered seriously and carefully. However, more research would be required to establish clearer evidence of the risks that foundation models may facilitate[76] — like the creation or deployment of chemical, biological, radiological, and nuclear weapons — before justifying broad restrictions on access to foundation models or the scientific discourse around them.

To be sure, we appreciate the reasonable desire to create conditions where researchers and foundation model developers spend sufficient time and are sufficiently motivated to evaluate increasingly advanced technologies, gather and — where appropriate — share insights to determine whether marginal indicators of these risks can be detected, and reduce them prior to such systems being widely shared. However, the government's role in aggressively intervening in the absence of particularized risk has traditionally been constrained in order to prevent arbitrary limitations on civil liberties or on the advancement of science.[77] Through that lens, the case for there being a substantial marginal risk from existing or imminent open models compared to other technologies has not yet been adequately made.

The Fragility of Safeguards in Open and Closed Models

One key factor to consider when examining the state of marginal risk is the robustness of guardrails once models are fine-tuned, whether the models are open or closed. As the AI EO highlighted when calling for this public consultation, "[w]hen the weights for a dual-use foundation model are widely available — such as when they are publicly posted on the Internet

---

current research is insufficient to effectively characterize the marginal risk of open foundation models relative to pre-existing technologies."

[75] *Letter to Laurie Locascio*, U.S. Committee on Science, Space, and Technology (Dec. 14, 2023), https://republicans-science.house.gov/_cache/files/8/a/8a9f893d-858a-419f-9904-52163f22be71/191E586AF744B32E6831A248CD7F4D41.2023-12-14-aisi-scientific-merit-final-signed.pdf [perma.cc/9XN8-32VZ]. (omitting internal citations) ("Organizations routinely point to significant speculative benefits or risks of AI systems but fail to provide evidence of their claims, produce nonreproducible research, hide behind secrecy, use evaluation methods that lack construct validity, or cite research that has failed to go through robust review processes, such as academic peer review."); See also Shazeda Ahmed et al., "Field-building and the Epistemic Culture of AI Safety," *First Monday (Forthcoming)*, 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4641526 [perma.cc/9YAR-W3YY].

[76] Bommasani et al., 2023, *supra* note 13 (comparing papers and noting key gaps where additional research is necessary).

[77] See especially discussion of the First Amendment in Part III.

— there can be substantial benefits to innovation, but also substantial security risks, such as the removal of safeguards within the model."[78]

However, the same research that demonstrated the fragility of safeguards in open foundation models (in this case Llama-2) also discovered the very same weaknesses in the guardrails provided by closed foundation models (in this case GPT-3.5 Turbo fine-tuned via its API).[79] There, researchers fine-tuned both kinds of models on a handful (<100) of harmful instructions and responses. Results demonstrated that this procedure largely undermined existing safeguards in both kinds of models, enabling the fine-tuned models to produce harmful outputs across many categories that model safeguards would have otherwise prevented (e.g., illegal activity, hateful content, physical harm, adult content). Yet those were not the most worrisome findings. The researchers also discovered, using common datasets to simulate scenarios where downstream actors might attempt to fine-tune models for a specific purpose, that even such benign fine-tuning led to a notable increase in harmful responses from GPT. Together, these results suggest that model guardrails are not robust against downstream modification in open foundation models or closed ones, even if downstream actors do not seek to intentionally circumvent protections. That calls into question the validity of the claim that this is a key differentiating concern between open and closed foundation models.

Additional research using somewhat different methods has also led to substantial demonstrations of the limitations of guardrails deployed by closed foundation models. For example:

- One study showed that fine-tuning GPT-4 (via API) on 340 examples can successfully undermine alignment at a rate of 95%. In other words, even more advanced models are not necessarily resilient against the types of attacks that are effective on less advanced models.[80]
- Another study identified three major categories of jailbreak prompts for both GPT-3.5 and GPT-4 and demonstrated that these prompts successfully undermined models' defenses at a rate of 74.6%.[81]
- A study also found that GPT-4 could be used to automate the discovery of new prompts to jailbreak other language models, including Claude 2, Vicuna, and itself.[82]
- Yet another study demonstrated how low-resource languages (e.g., Zulu, Hmong) are a mechanism for jailbreaking GPT-4 using just prompt-level access. They evaluated jailbreaks using 12 languages of varying resource levels by taking a harmful prompt in English, translating it into another language using Google Translate, feeding it into GPT-4, and then translating that output back into English. Combining low-resource

[78] AI EO, *supra* note 2.
[79] Qi et al., 2024, *supra* note 55.
[80] Qiusi Zhan et al., "Removing RLHF Protections in GPT-4 via Fine-Tuning," *arXiv*, November 9, 2023, https://arxiv.org/abs/2311.05553. [perma.cc/NY9X-SDVJ]
[81] Yi Liu et al., "Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study," *arXiv*, May 23, 2023, https://arxiv.org/abs/2305.13860. [perma.cc/FU4Y-76ZY]
[82] Rusheb Shah et al., "Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation," arXiv.org, November 6, 2023, https://arxiv.org/abs/2311.03348. [perma.cc/HM8J-BS7J]

languages allowed researchers to jailbreak GPT-4 79% of the time.[83] This result is particularly worrisome not only because it is a technique that can be used by attackers, but also because it reflects how safety filters will work substantially less well for regular users who speak low-resource languages.[84]

Research with similar results abound,[85] and more can be expected in the future. We can also expect more public examples of the fragility of guardrails, as we see more public controversies involving content generated by closed models. For example, the recent spate of fake Taylor Swift sexual content flooding social networks was not caused by open models; the images were generated by Microsoft's Illustrator using OpenAI's Dall-E.[86] Similarly, the controversy over robocalls using a synthesized Joe Biden voice started with a closed voice synthesis tool from ElevenLabs.[87] And AI-generated sexualized content regarding children was recently found on Shutterstock, reportedly using Shutterstock's image generator which also runs on a combination of closed models, OpenAI's Dall-E and LG's EXAONE.[88]

These examples highlight the ways that open and closed models face many of the same vulnerabilities — but also highlight one way in which open models do differ from closed ones. In the above examples, Microsoft was able to tweak its filters to catch the exploits based on misspellings that were used to generate the Taylor Swift images; ElevenLabs was able to block the user who created the Biden robocalls, and hopefully is developing audio fingerprinting blacklists to prevent similar harmful instances of impersonation of political figures; and Shutterstock was able to remove the content that had been created and posted on its site.[89]

---

[83] Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach, "Low-Resource Languages Jailbreak GPT-4," *arXiv*, October 3, 2023, https://arxiv.org/abs/2310.02446..
[84] Gabriel Nicholas, "Lost in Translation: Large Language Models in Non-English Content Analysis - Center for Democracy and Technology," *Center for Democracy and Technology* (blog), May 23, 2023, https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/. [perma.cc/5WK2-WXAK]
[85] See e.g., Zou et al., *supra* note 56 and Javier Rando et al., *Red-Teaming the Stable Diffusion Safety Filter*, *ML Safety Workshop NeurIPS 2022*, 2022, https://arxiv.org/abs/2210.04610 [perma.cc/2SW2-WBNV]. For a summary, see Nathan Lambert, "Undoing RLHF and the Brittleness of Safe LLMs," *Interconnects* (blog), October 18, 2023, https://www.interconnects.ai/p/undoing-rlhf [perma.cc/FU2M-KJMS].
[86] Carl Franzen, "Microsoft Adds New Restrictions to Designer AI Used to Make Taylor Swift Deepfakes," *VentureBeat*, January 29, 2024, https://venturebeat.com/business/microsoft-adds-new-restrictions-to-designer-ai-used-to-make-taylor-swift-deepfakes/. [perma.cc/RUQ4-A7KM]
[87] Margi Murphy, Rachel Metz, and Mark Bergen, "AI Startup ElevenLabs Bans Account Blamed for Biden Audio Deepfake," *Bloomberg*, January 26, 2024, https://www.bloomberg.com/news/articles/2024-01-26/ai-startup-elevenlabs-bans-account-blamed-for-biden-audio-deepfake. [perma.cc/C63Y-KN5L]
[88] Matt Growcoot, "Disturbing AI Images of Children Found for Sale on Shutterstock," *PetaPixel*, February 22, 2024, https://petapixel.com/2024/02/22/disturbing-ai-images-of-children-found-for-sale-on-shutterstock/. [perma.cc/U689-3B5W]
[89] Id.

By contrast, in the case of OFMs, the lack of a centralized provider means post-hoc enforcement action or central patching of vulnerabilities is not feasible. There is no way to rescind the publication of an open model once it is on the open web (notably true of all open web content already, even the most harmful), and no way for the publisher to consistently monitor or enforce against its users.[90] This structural difference is clearly relevant when considering safety threats from OFMs compared to closed foundation models. That said, reliance on closed model deployers to appropriately prioritize addressing harms in their models and to sufficiently enforce against misuse is likely to be imperfect, in the same way that online platforms can reduce prevalence of — but cannot entirely remove — conduct that violates their rules.

An assessment of marginal risk also must consider the risk of OFMs (and closed foundation models) compared to the existence of other, smaller models that are both already widely available and difficult to contain. Therefore, as we turn to evaluate different specific risks in turn, it is important to note that for all of them (other than the less well-defined "emergent risks" tied to computing on the frontier of current capability), smaller models that are specialized for the issuer at hand — such as for synthesizing DNA[91] or for writing the software code for hacking tools[92]— can create outputs that are equally if not more harmful than supposedly more capable tools.

Chemical, Biological, Radiological, and Nuclear Risks

No one has yet clearly demonstrated a marginal risk of bad actors misusing foundation models to facilitate the creation or deployment of chemical, biological, radiological, or nuclear (CBRN) weapons. That's because, as the Department of Justice concluded over twenty years ago, "anyone interested in manufacturing a bomb, dangerous weapon, or a weapon of mass destruction can easily obtain detailed instructions from readily accessible sources, such as legitimate reference books, the so-called underground press, and the Internet."[93]

---

[90] See Kapoor et al., 2024, *supra* note 13.

[91] See, e.g., Eric Nguyen et al., "Evo: DNA Foundation Modeling From Molecular to Genome Scale," *Arc Institute* (blog), February 27, 2024, https://arcinstitute.org/news/blog/evo [perma.cc/WK4R-VG6S] (describing a specialized model for generating DNA sequence that has only 7 billion parameters); Sara R. Carter et al., "The Convergence of Artificial Intelligence and the Life Sciences," *The Nuclear Threat Initiative*, October 30, 2023, https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences [perma.cc/X4XT-53GZ] (discussing how specialized biodesign tools, unlike current or imminent LLMs, may be able to synthesize toxins and pathogens that do not occur in nature and may even be more harmful than natural agents).

[92] E.g., WormGPT and other LLMs used by cyberattackers were built on GPT-J, a model comparable in size to GPT-2: Polra Victor Falade, "Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, October 3, 2023, 185-98, https://doi.org/10.32628/cseit2390533. [perma.cc/W7F6-MPQE]

[93] Report on the Availability of Bombmaking Information, the Extent to Which Its Dissemination Is Controlled by Federal Law, and the Extent to Which Such Dissemination May Be Subject to Regulation Consistent with the First Amendment to the United States Constitution: Prepared by the United States Department of Justice as Required by Section 709(a) of the Antiterrorism and Effective Death Penalty Act of 1996 (April 1997), quoted in CRS Report for Congress: Bomb-Making Online: Explosives, Free

What was true then remains true now, as demonstrated by two very recent studies intended to assess the biorisk threat posed by foundation models. Each concluded that the information provided by foundation models that may be useful in creating or deploying a bioweapon was essentially similar to what one could obtain with access to the internet.

In particular, a study by RAND conducted an exercise where research teams role-played as malign nonstate actors tasked with planning a biological attack; some were given access to an LLM, others only to the internet. The authors "found no statistically significant difference in the viability of plans generated with or without LLM assistance…. [O]utputs generally mirror[ed] information readily available on the internet."[94] OpenAI conducted a similar study around the same time to evaluate the biorisk from its GPT-4 model, again dividing researchers into role-playing teams with and without access to the model. They found only "mild uplifts" in the performance of the LLM-assisted teams, "too small to be statistically significant," and across only two of the five metrics that were tested.[95]

Meanwhile, previous papers claiming the existence of a severe AI-driven biorisk from foundation models suffer from the defects raised by the House Science Committee, in particular cross-citing to other papers for the general proposition of such risk but without substantiating evidence, or citing to sources that do not directly support that proposition.[96]

There has been no comparable, publicly available research into nuclear and chemical as opposed to biological risks from foundation models, and as such, no public evidence exists at this time of marginal risk in those areas. It would be reasonable to assume that the results of such testing in these domains would be similar to that of biorisk: the foundation model provides information that may be helpful, but reflects only information that it learned from the internet that bad actors can already access.

Even if foundation models did create a significant marginal risk of increasing the practical knowledge of bad actors to develop CBRN weaponry compared to the internet, it is not clear that such knowledge would increase the marginal risk of an actual attack (or that restricting access to OFMs would reduce that risk), because the primary bar to such attacks appears to be

---

Speech, Criminal Law and the Internet, Sep. 8, 2003, at p. 7, available at https://www.everycrsreport.com/files/20030908_RL32074_fcbf5a7d23f14b3350d4c2d81465aaaf7bcd299d.pdf. [perma.cc/JC27-FKRQ]

[94] Christopher A. Mouton, Caleb Lucas, and Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study," *RAND*, January 25, 2024, https://www.rand.org/pubs/research_reports/RRA2977-2.html. [perma.cc/7J76-JULN]

[95] Tejal Patwardhan et al., "Building an Early Warning System for LLM-aided Biological Threat Creation," *OpenAI*, January 21, 2024, https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation. [perma.cc/HZB4-G7NQ]

[96] "Propaganda or Science: Open Source AI and Bioterrorism Risk," 1A3ORN (blog), November 2, 2023, https://1a3orn.com/sub/essays-propaganda-or-science.html (examining and critiquing all of the biorisk-relevant citations in an influential policy paper arguing for new restrictions on OFMs); see also Kapoor et al., 2024, *supra* note 13. [perma.cc/S5A2-ZZ87]

material and logistical, rather than informational.[97] This was the conclusion of research from the Future of Life Institute, one of the primary organizations concerned with catastrophic risks of AI. A paper by one of the Institute's research analysts, who is also a senior researcher at Johns Hopkins specializing in biorisk, concluded that due to the practical requirements of developing and deploying a biological threat agent, "(m)alevolent low resourced actors and benevolent or accidental actors regardless of resource level are revealed as being unable to produce such [a biological] agent."[98] This is regardless of what information is available to those actors because the primary barrier to those actors is not lack of information. Furthermore, to the extent the government seeks to prevent facts about science from becoming less practically obscure and more easily findable, its interventions may violate the First Amendment as discussed in Part III.

The above considerations highlight the mismatch between the types of policy interventions being proposed to try to limit access to knowledge through OFMs, and the practical steps actually needed to secure the physical assets and facilities necessary to create, test, and deploy a bioweapon. Or, "[p]ut differently, which of the following may be more likely by 2024: more (a) open-source models, (b) laboratories capable of manufacturing pathogens, or (c) suppliers of required raw materials? If the answer is (a), the focus on (b) and (c) may provide more effective mechanisms of control."[99] This is all the more true considering the likely growth not only of OFMs but smaller and harder-to-police models that are specialized in biology.

Based on similar reasoning, the report of a convening of senior experts hosted by the Rockefeller Foundation examining "Biosecurity in the Age of AI" contained six policy proposals, only one of which was focused on attempting to guardrail the use of LLMs to prevent access to biothreat-relevant information. The rest were focused on much more practical and likely effective measures targeted at safeguarding the digital-to-physical frontier (e.g. through mandatory screening around DNA synthesis), investment in early detection and response, development of new lab safety norms, etc.[100]

Particularly considering the experience of the COVID-19 pandemic, investments in detection and response in particular could assist society in countering both AI-derived and natural bioagents. AI is and will be a key tool in that toolbox. As the nonprofit, nonpartisan Nuclear

---

[97] See e.g., Louise Matsakis, "Why AI-assisted Bioterrorism Became a Top Concern for OpenAI and Anthropic," *Semafor*, November 15, 2023, https://www.semafor.com/article/11/15/2023/ai-assisted-bioterrorism-is-top-concern-for-openai-and-anthropic. [perma.cc/C7V4-S57E]

[98] Michael Montague, "Towards a Grand Unified Threat Model of Biotechnology," *PhilSci-Archive*, 2023, https://philsci-archive.pitt.edu/22539/. [perma.cc/L2JJ-D859]

[99] Neel Guha et al., "AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing," *George Washington Law Review (Forthcoming)*, November 15, 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4634443. [perma.cc/9N6H-BHBA]

[100] Mark Dybul, "Biosecurity in the Age of AI Chairperson's Statement," *Helena at the Rockefeller Foundation's Bellagio Center*, June 2023, https://938f895d-7ac1-45ec-bb16-1201cbbc00ae.usrfiles.com/ugd/938f89_74d6e163774a4691ae8aa0d38e98304f.pdf [perma.cc/2JK9-T77U]. See also Rishi Bommasani et al., 2023, *supra* note 13. ("As with many other threat vectors, the best policy choke points may hence lie downstream. For example, the U.S. AI Executive Order aims to strengthen customer screening for purchasers of biological sequences.")

Threat Initiative highlighted in a recent paper, after concluding (like other studies) that general purpose LLMs are unlikely to generate toxin or pathogen designs that are not already described in the public literature: "AI-bio capabilities will also benefit society and bolster biosecurity and pandemic preparedness. In addition to broadly enabling scientific progress, AI models are already aiding pathogen biosurveillance systems, the development of medical countermeasures, and other aspects of pandemic preparedness and response."[101]

Attempting to restrict open access to biological capabilities could threaten those benefits — and by disrupting the offense-defense balance, may heighten rather than reduce biorisk. "[I]n the long run, the biosecurity solution to biotechnology is more biotechnology. Indeed, biosecurity policies that slow the adoption and advance of biotechnology artificially preserve and prolong a period of relative vulnerability in which defensive uses of biotechnology have yet to fully dominate the security equation."[102] The next section will discuss a similar dynamic in regard to cybersecurity threats.

Cybersecurity Risks

The Executive Order in its definition of a "dual-use foundation model" also highlighted concern that such models may "enabl[e] powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks." However, as with CBRN threats, this threat is, so far, under-evidenced. In fact, Microsoft and OpenAI have published the results of a study on offensive uses of the LLMs they monitor and found "incremental" changes in "behaviors consistent with attackers using AI as another productivity tool on the offensive landscape," but did not yet observe "particularly novel or unique AI-enabled attack or abuse techniques resulting from threat actors' usage of AI."[103]

Of course, newer models will likely offer more powerful opportunities for creating tools to help discover and exploit vulnerabilities in other systems. Yet whether those new capabilities are a mere incremental change or a large step change, those same capabilities will be available to defenders as well. Defenders will be able to discover the same vulnerabilities as the attackers, and work to patch them. Defenders will be assisted in their coding by LLMs the same as attackers. Defenders will be able to work to counter LLM-generated phishing messages with LLM-based detection of the same, much as we have developed automated tools that catch most human-generated spam. It is because of such benefits in the context of regular software — and likely First Amendment concerns, see Part III — that current export controls on "cybersecurity software" do not apply to publication of open source software.[104]

---

[101] Carter et al., 2023, *supra* note 91.
[102] Id.
[103] Microsoft Threat Intelligence, "Staying Ahead of Threat Actors in the Age of AI," *Microsoft Security Blog* (blog), February 14, 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai. [perma.cc/7F5R-U7ZG].
[104] See Department of Commerce, Information Security Controls: Cybersecurity Items, 86 Fed. Reg. 58205,58207 (Oct. 21, 2021); https://www.govinfo.gov/content/pkg/FR-2021-10-21/pdf/2021-22774.pdf [perma.cc/AWX5-DWY6] ("BIS does not intend this note to require any additional compliance measures

This is not to say that open sourcing powerful foundation models will certainly help defenders as much or more than attackers, nor do we intend to make light of the risks. There are a number of ways that open source software code and open models with weights are quite different artifacts, such that the cybersecurity risks and benefits of open models may differ somewhat. However, policymakers also cannot assume that they will certainly help attackers more, considering the significant cybersecurity benefits that openness in software and data have previously demonstrated. Once again, more research and a fuller record demonstrating the likelihood of such risk is necessary to justify broad restrictions on general purpose AI tools.

Emergent Risks

Although more research is needed around CBRN and cyber threats, nowhere is more and better articulation and proof of risk needed than in the realm of what we will call "emergent" risks. These are broader, longer-term risks about AI models going "rogue" — evading human control through deception, escaping their servers and self-proliferating, and/or deliberately acting of their own accord against the aims of humans.

The idea of rogue AI is a naturally worrisome one, and a common trope in science fiction for as long as we've conceived of artificial intelligence. However, these risks are considered speculative even by many of the experts who raise them.[105] For example, a highly cited paper on the topic — the same paper from which the AI EO apparently took its language focused on foundation models "evading human control through means of deception and obfuscation"[106] — footnoted its concern around this risk as follows:

> *If* future AI systems develop the ability and the propensity to deceive their users, controlling their behavior *could* be extremely challenging. *Though it is unclear whether models will trend in that direction*, it seems rash to dismiss the possibility and some

---

beyond what is otherwise required by the EAR. ''Software'' and ''technology'' ''published'' in the public domain and meeting the requirements of § 734.7 of the EAR are not subject to the EAR); see also EAR § 734.7 ("'unclassified 'technology' or 'software' is 'published,' and is thus not 'technology' or 'software' subject to the EAR, when it has been made available to the public without restrictions upon its further dissemination such as through any of the following….") and "Understanding US Export Controls With Open Source Projects," *The Linux Foundation*, July 2020, https://www.linuxfoundation.org/resources/publications/understanding-us-export-controls-with-open-source-projects (summarizing application of EAR to open source) [perma.cc/8LY3-5T47].

[105] "While currently deployed foundation models pose risks, they do not yet appear to possess dangerous capabilities that pose severe risks to public safety as we have defined them. Given both our inability to reliably predict what models will have sufficiently dangerous capabilities and the already significant capabilities today's models possess, it would be prudent for regulators to assume that next-generation state-of-the-art foundation models *could* possess advanced enough capabilities to warrant regulation." (emphasis in original, internal citations omitted). Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," *arXiv*, November 7, 2023, https://arxiv.org/abs/2307.03718 [perma.cc/X3CR-P5LH].

[106] Cf. id., listing as a key threat after CBRN and cyber threats the "[e]vading [of] human control through means of deception and obfuscation."

> argue that it might be the default outcome of current training paradigms. [Emphasis added.][107]

We would agree that these concerns should not be dismissed out of hand — even if the risk were very low, the threat to the public and humanity more broadly if that risk were to come to pass may be quite high. However, the main arguments for there being a risk of rogue AI originated with philosophers focused on catastrophic risk, years before the rise of LLMs and other generative AI technologies, and have not changed substantially since.[108] This suggests that these concerns are not primarily based on specific technical developments but reflect more philosophical extrapolations about how a hypothetical artificial mind might behave. And papers arguing that these catastrophic risks exist or are imminent typically do not present specific factual evidence, instead theorizing generally about the possibility of each kind of risk, sometimes with supporting anecdotes or illustrative potential scenarios.[109] Furthermore, existential risk scholars often prioritize outcomes based on the magnitude (positive or negative) of their consequences and their probability of occurring; however, these scholars' notions of both probability[110] as well as magnitude[111] are highly subjective. As a result, even when catastrophic risks are assigned a discrete numeric value, this value is better interpreted as a qualitative belief than as a precise quantitative estimate.

Those concerned about catastrophic emergent risks often point to the rapid increase in LLM's general capabilities as they are trained with more data and compute, to argue that we should anticipate sharp and unpredictable changes in those capabilities over time.[112] Indeed, one paper has suggested based on certain metrics that we have already seen such sharp, unpredictable changes in capability that may even demonstrate "the sparks of artificial general intelligence" and the capability to reason beyond the model's training data.[113] However, there are a number of reasons that skepticism around these assertions may be warranted absent more evidence.

---

[107] Id. (citations omitted).
[108] e.g., Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2015); Toby Ord, *The Precipice: Existential Risk and the Future of Humanity*. (Hachette Books, 2020), https://theprecipice.com/; For further analysis on this point see Ahmed et al., 2024, *supra* note 75. [perma.cc/K275-MTQ7].
[109] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," *arXiv*, October 9, 2023, https://arxiv.org/abs/2306.12001 [perma.cc/CPF3-R46A]; Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives," *Center for the Governance of AI*, September 29, 2023, https://arxiv.org/abs/2311.09227 [perma.cc/E2SG-7BU7]; Jeremie Harris, Edouard Harris, and Mark Beall, "Survey of AI Technologies and AI R&D Trajectories," *Gladstone AI*, November 3, 2023, https://assets-global.website-files.com/62c4cf7322be8ea59c904399/65e83959fd414a488a4fa9a5_Gladstone%20Survey%20of%20AI.pdf [perma.cc/FC7U-AV3W].
[110] Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology* 9 (2002), https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c. [perma.cc/Q95X-J5M5].
[111] Owen Cotton-Barratt and Toby Ord, "Existential Risk and Existential Hope: Definitions," *Future of Humanity Institute*, 2015, https://amirrorclear.net/files/existential-risk-and-existential-hope.pdf. [perma.cc/E493-E26Q].
[112] Seger at al., 2023, *supra* note 109.
[113] Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments With GPT-4," *arXiv*, March 22, 2023, https://arxiv.org/abs/2303.12712. [perma.cc/S4KM-NZ6M].

First, additional research has argued that the appearance of such discontinuous jumps depends heavily on the metrics used, and when using different metrics, the state of improvement — although very fast — follows a continuous curve.[114] Second, there is also the challenge of measuring improvements in model capability using tests that may have been in the training data of the model in the first place, such that those scores may not necessarily indicate any improvement in performance, much less reasoning.[115] Third, to the extent the general capabilities of foundation models increase, those capabilities will also be available to human users, providing a countervailing benefit that must be considered, including in how it could help balance the threat. Finally, it is again important to consider the practicalities of a rogue AI actually causing substantial catastrophic or even existential harm to humanity. As various commentators have highlighted, it is still very unclear how an AI model would gain control over the many physical assets it would likely need to create such a risk.[116]

Global Competition and Security Risks

Another concern sometimes raised, although not explicitly in the AI EO, is that open sourcing foundation models will assist China in competing with us economically or militarily. This is certainly true to the extent that open sourcing foundation models will give some new advantage to anyone seeking to build AI functionality without relying on a handful of companies offering closed models. However, it is also true that many of the largest OFMs on the leaderboard of the Hugging Face platform are of Chinese origin,[117] and although there is some reporting that one

---

[114] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, *Are Emergent Abilities of Large Language Models a Mirage?, Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023, https://arxiv.org/abs/2304.15004. [perma.cc/3S53-68L2]

[115] Contamination of training data: Tom Brown et al., Language Models Are Few-Shot Learners, *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html [perma.cc/CMY2-BXYS]; Jason Wei et al., "Finetuned Language Models Are Zero-Shot Learners," *arXiv*, September 3, 2021, https://arxiv.org/abs/2109.01652 [perma.cc/3UGY-84YA]; Simone Balloccu et al., Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs, Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, https://aclanthology.org/2024.eacl-long.5 [perma.cc/UHH6-5869]. Effects of contamination on performance: Changmao Li and Jeffrey Flanigan, "Task Contamination: Language Models May Not Be Few-Shot Anymore," *arXiv*, December 26, 2023, https://arxiv.org/abs/2312.16337 [perma.cc/X4HT-DWU5]. Federico Ranaldi et al., "Investigating the Impact of Data Contamination of Large Language Models in Text-to-SQL Translation," *arXiv,* February 12, 2024, https://arxiv.org/abs/2402.08100 [perma.cc/XZ7Q-E5HH].

[116] Timothy B Lee, "The AI Safety Debate Is Focusing on the Wrong Threats," *Understanding AI* (blog), May 9, 2023, https://www.understandingai.org/p/why-im-not-worried-about-ai-taking. [perma.cc/HZ9Y-S4ED].

[117] Sorting the Open LLM Leaderboard on Hugging Face at https://hf.co/spaces/HuggingFaceH4/open_llm_leaderboard [perma.cc/ZAZ8-ASVM] based on size and looking only at models greater than 35 billion parameters, there are multiple models originating from China including Qwen v1(and v1.5), Yi, and DeepSeek, along with Falcon (UAE) and Mixtral (France), but the only US-origin model in that category is Llama-2.

Chinese startup used a variant of Meta's Llama-2 architecture in training its LLM,[118] we have not yet seen examples of major Chinese OFMs using Llama-2 weights.

Meanwhile, we *have* seen a flourishing of OFMs of international origin that China can and will have access to regardless of US policy, including Falcon (United Arab Emirates),[119] Baichuan (China),[120] BLOOM (global),[121] Mixtral (France),[122] and Stable Beluga (England).[123]. Therefore it is unclear how artificially constraining international access to US-origin OFMs would substantially alter the course of international OFM development other than to slow it down (for lack of our contributions) and potentially give other international OFM developers a better chance to dominate the OFM development community and the foundation model market with their offerings while US-origin offerings are delayed.

Of course, as with all of the other national security risks above, our information may be, and in some cases certainly is, incomplete because of our lack of access to classified information. But the evidence currently available to the public suggests that heavy-handed interventions to restrict OFM exports are unlikely to meet their intended goals.

Content Risks

Although not the focus of this proceeding, it would be remiss not to highlight that there are several categories of serious risks from foundation models — both open and closed — that are not speculative but are already being observed. These include the proliferation of AI-generated, photorealistic child sexual abuse imagery (CSAM) and nonconsensual intimate imagery (NCII), misinformation and disinformation, and fraudulent content such as phishing emails or voice-cloning.[124]

---

[118] Paul Mozur, John Liu, and Cade Metz, "China's Rush to Dominate A.I. Comes With a Twist: It Depends on U.S. Technology," *The New York Times*, February 21, 2024, https://www.nytimes.com/2024/02/21/technology/china-united-states-artificial-intelligence.html [perma.cc/9M93-NJP6]; but see Hailey Schoelkopf, Aviya Skowron, and Stella Biderman, "Yi-34B, Llama 2, and Common Practices in LLM Training: A Fact Check of the New York Times," *EleutherAI Blog* (blog), March 26, 2024, https://blog.eleuther.ai/nyt-yi-34b-response [perma.cc/6JM5-EZ75] (explaining why re-use of LLM architecture is unremarkable because all modern LLMs use a similar architecture).

[119] Ebtesam Almazrouei et al., "The Falcon Series of Open Language Models," *arXiv*, November 28, 2023, https://arxiv.org/abs/2311.16867. [perma.cc/RQN6-PUHG]

[120] Aiyuan Yang et al., "Baichuan 2: Open Large-scale Language Models," *arXiv*, September 19, 2023, https://arxiv.org/abs/2309.10305. [perma.cc/9AJJ-GZW6]

[121] BigScience Workshop et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," *arXiv*, November 9, 2022, https://arxiv.org/abs/2211.05100. [perma.cc/D9LQ-GDEK]

[122] Albert Q. Jiang et al., "Mixtral of Experts," *arXiv*, January 8, 2024, https://arxiv.org/abs/2401.04088. [perma.cc/83VC-CEWS].

[123] Stability AI, "Meet Stable Beluga 1 and Stable Beluga 2, Our Large and Mighty Instruction Fine-Tuned Language Models," *Stability AI*, November 8, 2023, https://stability.ai/news/stable-beluga-large-instruction-fine-tuned-models. [perma.cc/56MZ-JYX3]

[124] Sayash Kapoor and Arvind Narayanan, "How to Prepare for the Deluge of Generative AI on Social Media," Knight First Amendment Institute, June 16, 2023, https://knightcolumbia.org/content/how-to-prepare-for-the-deluge-of-generative-ai-on-social-media. [perma.cc/G7YC-R3GL].

These are all serious policy challenges, and how best to address these very real harms — both in the context of open and closed systems — is still unclear and subject to extensive debate elsewhere, including in Congress. Therefore they are not addressed in depth here. However, it is worth highlighting some key considerations that have come up in the context of the previously discussed risks, which are also relevant to considering how to address policy in regard to these types of content harms.

First, as with previous risks, the ability to effectively address these content issues at the level of a foundation model is currently unclear, considering how fragile the safeguards of both closed and open models are against even well-intentioned fine-tuning as well as adversarial attacks. Or, as two Princeton researchers put it in a recent essay, "safety is not a model property,"[125] at least not in terms of current foundation model architectures.

Furthermore, for a number of content-related risks the marginal risk between open and closed models is currently unclear, not only because of equally fragile guardrails but also because some of these objectionable forms of content such as mis- and dis-information were already very cheap to produce,[126] and do not require capacity to produce synthetic content at all.[127] It is certainly possible that open models may ultimately generate more objectionable material than closed, presuming some level of effective enforcement of terms of use by closed model providers. However, even assuming that is the case, OFMs may not present a marginal risk as compared to smaller, specialized open models that it likely will not be possible to interdict and that may pose an equal or greater risk of creating harmful content such as CSAM.

Therefore, just as a focus on hardening attack surfaces such as DNA sequencing labs makes sense in the biorisk context, so too may a focus on stemming harmful content types at their distribution chokepoints, such as social networks.[128] However, and as will be discussed in Part III, the government must in all its efforts ensure compliance with the protections of the First Amendment, noting that several of the content categories discussed are or may be protected speech depending on the facts.

---

[125] Arvind Narayanan and Sayash Kapoor, "AI Safety Is Not a Model Property," *AI Snake Oil* (blog), March 12, 2024, https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property. [perma.cc/AGB9-8WL6]

[126] Felix M. Simon, Sacha Altay, and Hugo Mercier, "Misinformation Reloaded? Fears About the Impact of Generative AI on Misinformation Are Overblown," *Misinformation Review, Harvard Kennedy Review*, October 18, 2023, https://doi.org/10.37016/mr-2020-127; Kapoor & Narayanan, 2023, *supra* note 124. [perma.cc/3L7S-8QNB]

[127] Lisa Fazio "Out-of-context photos are a powerful low-tech form of misinformation," PBS News Hour, February 18, 2020, https://www.pbs.org/newshour/science/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation. [perma.cc/SDM2-PHG7]

[128] See Bommasani, et al., 2023, *supra* note 13: "[T]he key bottleneck for effective influence operations is not disinformation generation but disinformation dissemination: Online platforms that control the reach of content are better targets for policy intervention." See also Josh A. Goldstein et al., "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," *arXiv*, January 10, 2023, https://arxiv.org/abs/2301.04246 [perma.cc/DF2P-RMN9]; Richard L. Hansen, *Cheap Speech How Disinformation Poisons Our Politics — and How to Cure It* (Yale University Press, 2022), https://yalebooks.yale.edu/book/9780300274097/cheap-speech/ [perma.cc/FXM9-WSNB].

Civil Rights Risks

As with content risks, harms to civil rights from use of AI models and systems are already apparent. With AI models of all kinds, research has time and again demonstrated first-order harms of both allocation and representation from AI models, particularly when deployed in consequential contexts.

Civil rights-related harms from foundation models can manifest in several ways. If products based on foundation models (like chatbots) are used to directly make or materially contribute to decisions about people's economic or legal circumstances, such as using consumer chatbots to conduct employment screening or employee evaluations, embedded stereotypes can lead to arbitrary and disparate impact.[129] If foundation models are modified or integrated into downstream, context-specific use cases, undesirable characteristics of the foundation model such as embedded gender bias may persist into the downstream task.[130] Alternatively, downstream modification like contextual fine-tuning and product design can introduce biases even if they were successfully suppressed in the foundation models.[131]

If models reflect or amplify stereotypes in content generation even outside of consequential decisions, this can lead to stigmatization and the ossification of exclusionary norms.[132] And when communities are underrepresented in data that is used to train foundation models or are disproportionately subject to second-order effects like economic displacement or misuse of new tools to disenfranchise voters, the benefits and harms of this technology could continue to be distributed in a dramatically uneven fashion.

However, none of these harms is unique to OFMs, and in fact many such harms have already been identified in closed foundation models,[133] suggesting a lack of apparent marginal risk to civil rights compared to the harms caused by narrower and more widely deployed systems (which remain concerningly under-addressed). Moreover, research is mixed on the extent to

---

[129] Leon Yin, Davey Alba, and Leonardo Nicoletti, "OpenAI's GPT Is A Recruiter's Dream Tool. Tests Show There's Racial Bias," *Bloomberg*, March 7, 2024, https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination. [perma.cc/89AU-MYBG]

[130] Seungjae Shin et al., *Neutralizing Gender Bias in Word Embeddings With Latent Disentanglement and Counterfactual Generation*, *Findings of the Association for Computational Linguistics (EMNLP 2020)*, 2020, https://doi.org/10.18653/v1/2020.findings-emnlp.280. [perma.cc/VX6C-7R9K]

[131] Ryan Steed et al., "Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, January 1, 2022, https://doi.org/10.18653/v1/2022.acl-long.247. [perma.cc/N5NW-8TRS]

[132] Irene Solaiman et al., "Evaluating the Social Impact of Generative AI Systems in Systems and Society," *arXiv*, June 9, 2023, https://arxiv.org/abs/2306.05949. [perma.cc/YW3A-4LP9]

[133] "Study Assesses GPT-4's Potential to Perpetuate Racial, Gender Biases in Clinical Decision Making," *ScienceDaily*, December 18, 2023, https://www.sciencedaily.com/releases/2023/12/231218150939.htm [perma.cc/7J9N-R9PP]; James O'Donnell, "LLMs become more covertly racist with human intervention," *MIT Technology Review*, March 11, 2024, https://www.technologyreview.com/2024/03/11/1089683/llms-become-more-covertly-racist-with-human-intervention/ [perma.cc/7PRP-Y8ZS].

which intrinsic biases in foundation models correlate with bias in downstream tasks for which those foundation models play a role. Some research has found there to be no such correlation, raising fundamental questions about the measurement validity of existing — and well-intentioned — model evaluations seeking to measure bias at the foundation model layer and extrapolate those findings to real-world contexts.[134]

Some presume that central bias mitigation efforts and monitoring capacity will enable foundation model developers and hosts to robustly address intrinsic biases and intervene in circumstances that are particularly harmful to civil rights, but we worry that this assumption is highly optimistic.[135] Large technology companies have demonstrated reluctance or inability to proactively address the various ways harmful biases manifest across contexts, both within products and through enforcement actions, and we do not see strong evidence that AI developers — even if well-intentioned — will behave in a significantly different fashion.

It is important to note that like the associated concept of AI safety, fairness is not a model property:[136] research has shown that algorithms that appear to be fair in isolation do not necessarily combine into fair systems, and that apparently unfair models can still be combined in a way that leads to fairer systems.[137] And fairness is highly contextual: different use-cases may demand different definitions of fairness or civil rights compliance,[138] which a universal set of measurements or interventions at the foundation model layer may not be capable of achieving simultaneously.

Ultimately, the civil rights-related impacts of foundation models will depend heavily on the contexts of their deployment; for instance, foundation models used in the context of housing would be subject to the Fair Housing Act's prohibitions around steering homebuyers toward or away from certain neighborhoods, while foundational models used in the context of credit would be subject to Equal Credit Opportunity Act's fair lending requirements around both disparities in access to credit as well as explanations of adverse actions. Even in a circumstance where the most advanced models are subject to pre-market testing to reduce the most egregious civil

---

[134] Seraphina Goldfarb-Tarrant et al., *Intrinsic Bias Metrics Do Not Correlate With Application Bias*, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, https://doi.org/10.18653/v1/2021.acl-long.150. [perma.cc/9JAS-SHJX]

[135] Naomi Nix, "Big Tech Is Failing to Fight Election Lies, Civil Rights Groups Charge," *Washington Post*, October 27, 2022, https://www.washingtonpost.com/technology/2022/10/27/civil-rights-2022-midterms/ [perma.cc/45D5-GWC3]; OpenAI has also been criticized for failing to enforce its policies for third party tools in its GPS store: Kyle Wiggers, "OpenAI's Chatbot Store Is Filling up With Spam," *TechCrunch*, March 20, 2024, https://techcrunch.com/2024/03/20/openais-chatbot-store-is-filling-up-with-spam/ [perma.cc/6BG9-GG6H].

[136] Narayanan & Kapoor, 2024, *supra* note 125.

[137] Cynthia Dwork and Christina Ilvento, *Fairness Under Composition*, *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, 2019, https://doi.org/10.4230/LIPIcs.ITCS.2019.33. [perma.cc/FK9G-6H6X]

[138] doaa Abu Elyounes, "Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness," *Journal of Law, Technology and Policy*, 2019, https://doi.org/10.2139/ssrn.3478296. [perma.cc/55T9-UGZ6]

rights violations, subtle biases can manifest in the contexts where AI-powered systems are deployed in unpredictable and varied ways.

For this reason, robust enforcement of civil rights laws at the point of deployment will likely prove a critical lever of accountability for adverse civil rights impacts, while interventions at the foundation model layer may provide few, if any, guarantees that concrete civil rights harms will be avoided.[139] Even so, and given the variety of ways these harms will likely manifest and the difficulty of detecting all of them within the four corners of a foundation model, it is all the more important to help a broader community of researchers and context-specific experts gain visibility into these systems and their use cases. Unfortunately, platforms of all kinds, including foundation model providers, have been known to actively prevent the very types of research activities that can reveal these harms.[140] These sorts of dynamics make broader access to cutting edge versions of these models all the more important. If the same foundation model developers that might fail in protecting marginalized communities from harm are also in a position to prohibit research on their models, it will be far more difficult for third-party experts to help spot and prevent harms.

III. POLICY APPROACHES TO OPEN FOUNDATION MODELS

The AI EO highlights the Administration's interest in "potential voluntary, regulatory, and international mechanisms to manage the risks and maximize the benefits" of open foundation models. This section will focus on two issues: first, the issue of governmental support for the establishment of clear best practices and norms around responsible development and deployment of foundation models generally, and OFMs in particular; second, the issue of First Amendment limits on how far the government can go in *requiring* those practices and norms.

Creating an Infrastructure for Better Understanding of Model Risks

As discussed above, the evidence does not yet support a conclusion that OFMs currently create a material marginal risk in areas such as the creation or deployment of chemical, biological, radiological, and nuclear weapons. At the same time, we cannot rule out the possibility that OFMs at some point in the future may create such risks. The government should begin creating the mechanisms necessary to better assess and monitor whether some future model crosses that risk threshold.

A critical step along this path is already taken in the Executive Order by vesting responsibility in NIST to help establish clearer testing benchmarks for a range of foundation model risks. Continued strong and steady investment in convening and research to develop technically

[139] Note that if a foundation model developer intends to directly deploy their model for use, they should both anticipate and take action through policies and technical mitigations to prevent civil rights harms
[140] Nitasha Tiku, "Top AI Researchers Say OpenAI, Meta and More Hinder Independent Evaluations," *Washington Post*, March 5, 2024, https://www.washingtonpost.com/technology/2024/03/05/ai-research-letter-openai-meta-midjourney. [perma.cc/RRR7-KL2G]

feasible and effective testing norms is crucial at this early stage of the emerging AI safety field, when we still lack clear and consistent standards to apply and also do not yet have a large field of experts to develop and apply those standards, whether in-house at AI companies or through consultancies or auditing companies.[141]

This lack of clear norms is exacerbated by the fact that it is not yet clear what an appropriate AI audit consists of or what it should be testing for.[142] Even worse, we do not really know how effective the tests that we have are: there is no shortage of research calling into question whether emerging AI evaluation methodologies actually effectively measure risk or have a meaningful relationship to what happens when a model is released into society.[143] Therefore, as CDT recently urged NIST in another proceeding, it and other elements of the government focused on AI best practices should focus on promoting (both within and outside the government) foundational investments in the basic risk management processes that are needed to provide a stable groundwork for appropriate risk evaluation and mitigation for AI models of all kinds.[144] As we highlighted there, "a common set of concepts, approaches, and infrastructure for AI risk management [generally] is needed to lay the foundation for generative AI-specific analysis and intervention,"[145] including basic approaches to designing and judging the validity of different methods of testing and evaluation. That is because, as highlighted in a recent blog post from CDT, "trustworthy AI needs trustworthy measurement."[146]

Alongside the development of more reliable tests to better understand the risks an foundation model poses, NTIA should consider how the government can best obtain the information needed to monitor whether an OFM has crossed a threshold that now presents material marginal risks so that it can determine any appropriate responsive policy actions. In part, that may involve market surveillance activities designed to keep abreast of foundation model capabilities. Policymakers should also consider what forms of information sharing and transparency from developers of OFMs may be necessary, though as discussed below any

---

[141] Abeba Birhane et al., "AI Auditing: The Broken Bus on the Road to AI Accountability," *arXiv*, January 25, 2024, https://arxiv.org/abs/2401.14462. [perma.cc/Y9NS-RZZ5]
[142] Id.
[143] For a general discussion of reliability and validity in AI, see Winecoff & Bogen, 2024, *supra* note 63. For a critique of red teaming, see Sorelle Friedler et al., "AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability," *Data & Society*, October 23, 2023,https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability [perma.cc/DTP3-FP7S]; for a critique on the validity of technical safety approaches, see Narayanan & Kapoor, 2024, *supra* note 125; for a challenge to existing legal benchmarks, see Peter B. Henderson et al., "Rethinking Machine Learning Benchmarks in the Context of Professional Codes of Conduct," *Symposium on Computer Science and Law (CSLAW '24)*, 2024, https://doi.org/10.1145/3614407.3643708 [perma.cc/W2EB-B3U5]. For evidence of unreliability of model prompt safeguards, see Terry Yue Zhuo et al., "Red Teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity," *arXiv,* January 30, 2023, https://arxiv.org/abs/2301.12867 [perma.cc/5N3W-D6TU].
[144] Miranda Bogen, Gabriel Nicholas, and Amy Winecoff, "CDT Comments to NIST on Its Assignments Under the Executive Order Concerning Artificial Intelligence - Center for Democracy and Technology," *Center for Democracy and Technology*, February 2, 2024, https://cdt.org/insights/cdt-comments-to-nist-on-its-assignments. [perma.cc/7UY3-YYAQ]
[145] Id.
[146] Winecoff & Bogen, 2024, *supra* note 63.

compelled disclosures would be subject to First Amendment scrutiny and would need to be narrowly tailored and well-designed.

Promoting Safety Norms and Best Practices for Responsible Foundation Model Development and Release

Especially since the launch of GPT-3 and ChatGTP in 2022, there has been an enormous wave of activity from AI labs, the open source community, civil society, and policymakers seeking to establish clearer norms around how to develop, deploy, and use foundation models responsibly. We commend the Administration for securing significant voluntary commitments from many of the largest foundation model developers based on many of these initial practices.[147] The work kicked off by the AI EO and the upcoming OMB memo to agencies on responsible deployment of AI will also help to lay a firmer foundation for best practices in this area.[148]

Although the voluntary commitments mostly applied to larger closed model providers, we are beginning to see parallel norm development in the OFM space. For example, corporate developers like Google and Meta that have released OFMs have also been helping build norms around how to responsibly release open models, not only through publishing similar transparency artifacts about their models, as other labs do, but by releasing suites of materials and tools helpful to a deployer seeking to responsibly use the models. For example, with Llama-2, Meta released an extensive responsible user guide, walking through the key steps of mitigating risks in LLMs, and has begun releasing open source tools and evaluation datasets for security and content safety that deployers can use.[149] Upon the release of its Gemma open foundation models, Google similarly published a detailed Responsible Generative AI Toolkit with extensive advice, open source interpretability tooling, and methods for content filtering using AI classifiers.[150]

Crucially, both companies also have released versions of their models fine-tuned and red-teamed for usefulness and safety, for those who want to deploy them quickly with minimal customization — which is particularly important where their model licenses allow for commercial use and where the models they release may otherwise be put into service (inadvisedly) without

---

[147] The White House, "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments From Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," Press release, The White House, July 21, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/. [perma.cc/75SR-TGHV]

[148] Executive Office of the President, Office of Management and Budget, "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," Proposed Memorandum for the Heads of Executive Departments and Agencies, November 1, 2023, https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf. [perma.cc/HA22-6E33]

[149] Meta, "Purple Llama," n.d., https://llama.meta.com/purple-llama/ [perma.cc/2NZ8-PJKV].

[150] Google, "Responsible Generative AI Toolkit," n.d., https://ai.google.dev/responsible . [perma.cc/2LYP-F954]

adequate safeguards.[151] These and other highly capitalized AI companies should sharply increase investment in these sorts of transparency, safety, and accountability efforts and artifacts.

AI researchers from the academic and nonprofit worlds are also self-organizing to develop a wide range of resources for those seeking advice and tools for building OFMs responsibly. For example, the open AI engineering consortium MLCommons, which builds and maintains a wide range of test data sets and evaluation tools around accuracy, speed, and efficiency, is now developing new evaluations for safety issues and societal risks.[152] Meanwhile, a coalition of universities and non-profit labs have also developed the Foundation Model Cheat Sheet, a growing central repository of responsible development guidance and tools built by OFM developers for OFM developers.[153] As already mentioned, we are also seeing innovations in software licensing from both commercial and non-commercial players, as developers experiment with use restrictions in their AI licenses that can potentially support liability for or takedowns of noncompliant deployers,[154] while the Open Source Initiative is collaboratively developing its own new open source AI license.[155]

A particularly promising area of study is in the spectrum of release options between models that are not open at all to fully open source OFMs, with an OSI-compliant license with no use restrictions and open data/weights. As researcher Irene Solaiman was one of the first to highlight, between those two poles developers can make a lot of choices about when to release what components to whom in order to maximize safety and minimize risk (for example, allowing researcher access for testing prior to publication, or otherwise making the model available for testing in a controlled environment before release; holding back models with particularly risky capabilities until more extensively tested; etc.).[156]

Building on this work, a diverse coalition of experts including CDT were recently convened by Columbia University and Mozilla to develop a more comprehensive mapping of the variety of dimensions of openness available to publishers, including breaking down the various pros and cons of releasing different types of model components or transparency artifacts under different licenses or to different audiences.[157] We hope that a more specific parsing of these factors will

---

[151] Meta, 2022, *supra* note 4; "Gemma Terms of Use," Google AI, February 1, 2024, https://ai.google.dev/gemma/terms. [perma.cc/JYM9-8WGG]
[152] MLCommons, "MLCommons Announces the Formation of AI Safety Working Group," October 26, 2023, https://mlcommons.org/2023/10/mlcommons-announces-the-formation-of-ai-safety-working-group. [perma.cc/PB3P-NTQX]
[153] AI2 et al., "The Foundation Model Development Cheatsheet," n.d., https://fmcheatsheet.org/. [perma.cc/7SNC-YGT6]
[154] BigScience, 2022, *supra* note 4.
[155] Mia Lykou Lund, "Open Source AI Definition — Weekly Update Mar 18," *Open Source Initiative* (blog), March 18, 2024, https://opensource.org/blog/open-source-ai-definition-weekly-update-mar-18. [perma.cc/2XFE-Y3XH]
[156] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," *arXiv*, February 5, 2023, https://arxiv.org/abs/2302.04844. [perma.cc/9A34-Z48N]
[157] Ayah Bdeir and Camille Francois, "Introducing the Columbia Convening on Openness and AI," *The Mozilla Blog* (blog), March 6, 2024, https://blog.mozilla.org/en/mozilla/ai/introducing-columbia-convening-openness-and-ai

enable better policymaking, whether privately at the OFM publisher level or publicly through regulation or legislation.

Software and the First Amendment

The question of what are or should be best practices in responsible AI development is distinct from the question of what best practices the government can or should require by law. That is because potential regulation of OFMs may raise serious First Amendment questions. U.S. circuit courts have consistently held that the creation and publication of software code is expressive and is therefore protected by the First Amendment.[158] That conclusion likely applies to the code underlying OFMs and potentially to other model artifacts.

In *Junger v. Daly*, the Sixth Circuit Court of Appeals held that encryption software source code was speech protected by the First Amendment and that export controls prohibiting its publication to the internet triggered First Amendment scrutiny: "Because computer source code is an expressive means for the exchange of information and ideas about computer programming, we hold that it is protected by the First Amendment."[159]

The same argument would apply here to the extent the government is aiming to prevent the expression of scientific knowledge — whether about AI or generated by AI. Or, as the Ninth Circuit put it in another case considering the constitutionality of encryption export controls, *Bernstein v. U.S. Department of Justice*, in a passage worth quoting at length:

> [C]ryptographers use source code to express their scientific ideas in much the same way that mathematicians use equations or economists use graphs…. [M]athematicians and economists have adopted these modes of expression in order to facilitate the precise and rigorous expression of complex scientific ideas. Similarly, the undisputed record here makes it clear that cryptographers utilize source code in the same fashion. In light of these considerations, we conclude that encryption software, in its source code form and as employed by those in the field of cryptography, must be viewed as expressive for First Amendment purposes, and thus is entitled to the protections of the prior restraint doctrine. If the government required that mathematicians obtain a prepublication license prior to publishing material that included mathematical equations, we have no doubt that such a regime would be subject to scrutiny as a prior restraint.

[perma.cc/29WJ-HVJR]. Initial technical and policy memos from this process will be posted at https://research.mozilla.org/ on Wednesday, 3/27/23. See also similar research efforts such as, e.g., Matt White et al., "The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI," *arXiv*, March 20, 2024, https://arxiv.org/abs/2403.13784 [perma.cc/L8NH-WLFB]; Partnership on AI. "PAI's Guidance for Safe Foundation Model Deployment," March 14, 2024. https://partnershiponai.org/modeldeployment/ [perma.cc/Z4PZ-BZZC].

[158] See Bernstein v. U.S. Dept. of Justice, 192 F.3d 1308 (9th Cir. 1999), reh'g granted, opinion withdrawn, 192 F.3d 1308 (9th Cir. 1999) and Junger v. Daley, 209 F.3d 481 (6th Cir. 2000) (holding that source code can be expressive); see also Universal City Studios v. Corley, 273 F.3d 429 (2d Cir. 2001) (holding that both source code and object code can be expressive).

[159] Junger, 209 F.3d at 481.

While the *Bernstein* court relied on prior restraint doctrine, the *Junger* court instead applied intermediate scrutiny because it found the regulation targeted the functionality rather than the expressiveness of the code. It further found that the government had not met its First Amendment burden to demonstrate how the export control restrictions were narrowly drawn to address a specific problem with a tailored solution. “The government must demonstrate that the recited [national security] harms are real, not merely conjectural, and that the regulation will in fact alleviate these harms in a direct and material way.”[160] The *Junger* court ruled against the government even though it acknowledged that encryption software could enable malicious actors to hide their actions from government surveillance.[161]

Of course, legal doctrine can change — the existence of these rulings are not necessarily dispositive of how courts will rule now, especially when there are some notable factual differences. In particular and as highlighted previously, most OFM system components are made of software code, and therefore likely protected by the First Amendment under these and other precedents. Model weights, however, are not code but more akin to a very complex machine-readable database mapping the strength of connections between billions of “tokens” — in the case of LLMs, portions of words — read from a corpus of training data. Therefore, weights cannot be comprehended directly by people. This may prove to be an important distinction, since the previous courts considered it important that at least some computer scientists could read and comprehend software code.[162]

On the other hand, model weights are arguably more expressive than encryption software code, despite not being readable by human eyes. Weights are a mathematical object reflecting the characteristics of the vast amount of human language or imagery in its training data. By “reading” those weights with inference software, users can receive a vast range of helpful (or unhelpful) expressive content derived from those weights, which in turn could support their creative visions or educational pursuits or business endeavors or scientific exploration.[163]

---

[160] Id. at 485 (internal quotations omitted).
[161] Id.
[162] *Bernstein*, 192 F.3d at 1308 (“The distinguishing feature of source code is that it is meant to be read and understood by humans and that it can be used to express an idea or a method.”); *Junger*, 209 F.3d at 484 (“Particularly, a musical score cannot be read by the majority of the public but can be used as a means of communication among musicians. Likewise, computer source code, though unintelligible to many, is the preferred method of communication among computer programmers.”); *Corley*, 273 F.3d at 445-46 (“Mathematical formulae and musical scores are written in ‘code,’ i.e., symbolic notations not comprehensible to the uninitiated, and yet both are covered by the First Amendment. If someone chose to write a novel entirely in computer object code by using strings of 1's and 0's for each letter of each word, the resulting work would be no different for constitutional purposes than if it had been written in English. The "object code" version would be incomprehensible to readers outside the programming community (and tedious to read even for most within the community), but it would be no more incomprehensible than a work written in Sanskrit for those unversed in that language.”).
[163] Courts have recognized that listeners and readers have a 1st amendment right to receive speech.. *See* Kleindienst v. Mandel, 408 U.S. 753, 762-63 (1972) (“In a variety of contexts this Court has referred to a First Amendment right to ‘receive information and ideas’ . . . .”); Stanley v. Georgia, 394 U.S. 557, 564 (1969) (“It is now well established that the Constitution protects the right to receive information and ideas.”). Therefore, even if one does not count the developer as the speaker of generated outputs because they are somewhat stochastic, and one does not count the system as a speaker

Viewed in this manner, weights are not only expressive but uniquely so, and therefore especially warranting First Amendment protection.

The First Amendment is not absolute, however.[164] Regulation of speech protected by the First Amendment is possible when the appropriate standards are met. Consequently, as the NTIA considers the available policy and regulatory options with respect to OFMs, it should consider the constitutional implications of each option and recommend ways in which each option might be designed to maximize the likelihood of meeting requisite First Amendment standards, while still achieving the government's legitimate regulatory goals.

To assist NTIA in that endeavor, we briefly discuss some of the regulatory options often mentioned in reference to OFMs, and the First Amendment concerns they raise, in descending order from most to least serious constitutional questions. Based on this discussion, we offer some practical advice on how to avoid recommending policy solutions that courts are more likely to find violate the First Amendment.

Prior Restraints including Pre-Licensing

Any requirement that creators or distributors of OFMs obtain a license from a government entity (or a private entity designated by the government) prior to making their model weights widely available would likely be viewed by a reviewing court as a prior restraint on publication.[165]

Courts generally view prior restraints on publication with deep skepticism, including in circumstances related to the protection of national security, and with good reason.[166] More than chilling speech, it "freezes" speech, at least for a time, and has permanent and irreversible negative effects.[167] For that reason, prior restraints are presumptively unconstitutional and a heavy burden rests with the government to justify their necessity.

To the extent such restraints are put in place, they will almost certainly fail First Amendment scrutiny absent strong procedural safeguards to help counter the burden on speech, such as

---

because it is not a person, the First Amendment still may be implicated. See, e.g., Eugene Volokh, Mark A. Lemley, and Peter Henderson. "Freedom of Speech and AI Output," *Journal of Free Speech Law*, August 3, 2023, https://www.journaloffreespeechlaw.org/volokhlemleyhenderson.pdf (arguing for the First Amendment protection of AI outputs based on the users' right to receive). [perma.cc/9VZJ-EWLY]

[164] United States v. Stevens, 559 U.S. 460 (Apr. 2010); Kathleen Ann Ruane. Freedom of Speech and Press: Exceptions to the First Amendment, Cong. Research Serv. (Sept. 8, 2014) https://digital.library.unt.edu/ark:/67531/metadc462149/. [perma.cc/66ZF-EHN4]

[165] See New York Times Co. v. United States, 403 U.S. 713, 714 (1971) (injunction sought by United States against publication of the Pentagon Papers denied as an unconstitutional prior restraint on publication); Freedman v. Maryland, 380 U.S. 51, (1965) ("a noncriminal process which requires the prior submission of a film to a censor avoids constitutional infirmity only if it takes place under procedural safeguards"); Near v. Minnesota, 283 U.S. 697, 716 (1931).

[166] New York Times, 403 U.S. at 719 ("The word "security" is a broad, vague generality whose contours should not be invoked to abrogate the fundamental law embodied in the First Amendment. The guarding of military and diplomatic secrets at the expense of informed representative government provides no real security for our Republic.").

[167] Nebraska Press Association v. Stuart, 427 U.S. 539, 559 (1976).

clear and objective criteria to reduce the discretion of the licensor, clear time limits for a decision to be made, and an ability for prompt judicial review of negative determinations. The lack of such protections was the final nail in the coffin for the licensing scheme in *Bernstein*; the government should avoid such a result here.[168]

Transparency Requirements

Transparency regarding training data, fine-tuning efforts, input and output filtering, and other factors that help make a foundation model and the products it powers understandable is strongly desirable as a best practice.[169] However, transparency requirements imposed by the government in the non-commercial context are generally considered to be compelled speech and trigger First Amendment scrutiny.[170] Transparency requirements in the commercial context are also subject to the First Amendment, but generally courts apply a lower standard of scrutiny.[171] In any case, NTIA should take into account the applicable standard when designing any transparency-related regulatory recommendations.

The degree of constitutional concern raised by a particular transparency requirement will depend at least in part upon the scope of models subject to the requirement and the extent to which the requirements might burden or influence the editorial judgment of model developers.[172] For instance, transparency requirements applicable to all OFMs, regardless of context, purpose, or distributing entity, will more likely be subject to strict scrutiny, necessitating the government to meet the high standard of demonstrating the requirements are the least restrictive means of achieving a compelling government interest.[173] On the other hand, if the requirements apply only

---

[168] Alan Estevez, "Fireside Chat with Under Secretary Alan Estevez," Center for Security and Emerging Technology (CSET), Georgetown University, December 2023, https://www.youtube.com/watch?v=WClaOr4wZMM&t=4325s [perma.cc/AW88-Z6BP] ("We're talking about … large language models, we're having those discussions … I have a team … working on what's the answer.") See also Karen Hao, "The New AI Panic," *The Atlantic*, October 2023, https://www.theatlantic.com/technology/archive/2023/10/technology-exports-ai-programs-regulations-china/675605/ [https://perma.cc/DYW8-NJ4A] ("Commerce is considering a new blockade on a broad category of general-purpose AI programs, not just physical parts, according to people familiar with the matter.")

[169] Caitlin Vogus and Emma Llansó, "Report - Making Transparency Meaningful: A Framework for Policymakers," *Center for Democracy and Technology*, December 14, 2021, https://cdt.org/insights/report-making-transparency-meaningful-a-framework-for-policymakers/. [perma.cc/JA6B-7L8V]

[170] Riley v. National Federation of the Blind of North Carolina, Inc, 487 U.S. 781 (1988); Meese v. Keene, 481 U.S. 465 (1987).

[171] Zauderer v. Office of Disciplinary Counsel, 471 U.S. 626, (1985) (holding that the government may require the disclosure of purely factual information in the commercial context, as long as the requirement is reasonably related to the government's interest and not unduly burdensome).

[172] See Daphne Keller, "Platform Transparency and the First Amendment," *Social Science Research Network*, March 7, 2023, https://doi.org/10.2139/ssrn.4377578 [perma.cc/7BG5-64Q8]; Kathleen Ann Ruane. Freedom of Speech and Press: Exceptions to the First Amendment, Cong. Research Serv. (Sept. 8, 2014) https://digital.library.unt.edu/ark:/67531/metadc462149 [perma.cc/XP43-NDUN].

[173] See Riley, 487 U.S.at 796-97 ("There is certainly some difference between compelled speech and compelled silence, but in the context of protected speech, the difference is without constitutional significance, for thenFirst Amendment guarantees "freedom of speech," a term necessarily comprising the decision of both what to say and what not to say.")

to commercial or for-profit publishers of OFMs, more permissible standards of scrutiny may be applied, making it easier for the government to justify their imposition.[174]

The level of constitutional concern will also turn on whether transparency requirements encompass disclosures that would impact the editorial judgment of model developers and distributors. For example, requirements to conform to or disclose performance against benchmarks related to a model's expressive outputs, including toxic speech, hate speech, election disinformation, and scientific knowledge that is undesirable but that publishers have a right to distribute, would raise serious constitutional questions to the extent that courts conclude that the transparency requirements at issue would impact the editorial judgment of the speakers.[175]

Another factor relevant to constitutionality is the breadth of the audience of the required disclosure. In terms of assessing compelled speech, courts have in the past found that narrower compelled disclosures in the context of a particular proceeding with strong procedural protections can comply with the First Amendment.[176] And a final factor to consider is exactly when disclosure is required. Disclosure that is required contemporaneous with or after the publication of an OFM does not raise the specter of a prior restraint, while courts may see required disclosure to the government prior to publication as an attempt at an informal prior restraint and opportunity for the government to pressure against the publication of otherwise protected speech. On the flip side, to the extent that a transparency requirement is intended to help inform the government of significant risks resulting from an OFM so that it can take appropriate action, a pre-release transparency requirement could be justified as necessary to serve the government's legitimate interests. How a court might resolve those competing considerations is unclear and may depend on the particular facts.

For these reasons, NTIA should think carefully about the government interests advanced by transparency into OFMs with open model weights, closely tie any requirements to fulfilling interests separate and distinct from restricting OFMs from outputting protected expression or knowledge, and avoid requirements that might burden or influence editorial judgment, by

[174] See Netchoice, L.L.C. v. Paxton, 49 F.4th 439 (5th Cir. 2022) (cert. granted) (finding requirements for social media platforms to disclose an acceptable use policy and information about content and business practices likely did not violate the First Amendment); NetChoice, LLC v. AG, Fla., 34 F.4th 1196 (11th Cir. 2022) (cert. granted) (finding that requirements for social media "platforms to publish their standards, inform users about changes to their rules, provide users with view counts for their posts, and inform candidates about free advertising," are likely not unduly burdensome nor likely to chill platforms' speech" in violation of the First Amendment). See also, Volokh v. James, 656 F. Supp. 3d 431 (S.D. NY. 2023) (holding requirements for social media companies to create mechanisms from reporting "hateful conduct" and disclose their policies regarding how they will respond to complaints likely violates the First Amendment).

[175] See NetChoice, LLC v. AG, Fla., 34 F.4th 1196 (11th Cir. 2022) (cert. granted) (finding that requirements for social media platforms to provide public individual justifications for each content moderation decision likely did violate the First Amendment because it is overly burdensome and likely chills protected speech); but see Netchoice, L.L.C. v. Paxton, 49 F.4th 439 (5th Cir. 2022) (cert. granted) (finding similar requirements likely did not violate the First Amendment).

[176] Herbert v. Lando, 441 U.S. 153, 169 (1979) (allowing compelled disclosure of editorial decisionmaking under court supervision in a particular proceeding).

focusing any recommended transparency on the production of factual and uncontroversial information about the models.

Impact Assessment and Risk Management Requirements

Like transparency reporting, internal processes for assessing the impacts and managing the risks of AI systems are also a desirable best practice. However, when the government imposes requirements for engaging in such processes around First Amendment-protected activities like the creation and publication of expressive software (and here, arguably, expressive weights), care must be taken in designing them to reduce the risk that a court might conclude they unduly interfere with editorial judgments. For example, courts may find that certain decisions about what expressive content the model is allowed to output are editorial and that requirements to assess the risks posed by that content are subject to First Amendment scrutiny (though such requirements may in some circumstances withstand that scrutiny).[177] As with the earlier discussion of transparency requirements, NTIA should where possible tailor its recommendations to reduce or eliminate First Amendment concerns with assessment and risk management requirements, such as by clearly targeting such requirements at non-expressive, functional aspects of a model's development and performance.

Context-Specific Requirements

The government also has greater authority to impose stronger regulations on the deployment and use of OFMs in specific applications and contexts. For example, in situations where models are being used to make determinations regulated by existing civil rights laws, including those regarding eligibility for housing, employment, credit, or other economic opportunities, the government has broad discretion to take measures, including requiring transparency, auditing, and training data restrictions, that would ensure that the models are not discriminating against individuals on the basis of their membership in a protected class.[178] As already described previously when discussing civil rights, a focus on enforcement at the deployment level may be most effective practically and policy-wise; it would also mitigate First Amendment concerns.

Additionally, the government has greater leeway to impose more stringent requirements on the foundation models it seeks to use for its own purposes[179] (hence the importance of the

---

[177] See Netchoice v. Bonta, 2023 U.S. Dist. LEXIS 165500 (N.D. CA 2023) (finding that, applying intermediate scrutiny, a requirement for commercial web sites to conduct risk assessments related to potential harms to children posed by their services and detailed plans to address those risks likely violates the First Amendment because the requirement was not properly drawn to address the government's legitimate interests).

[178] *See* Christian Legal Soc'y Chapter of the Univ. of Cal., Hastings Coll. of the Law v. Martinez, 561 U.S. 661, 694-95 (2010) (finding non-discrimination requirements to be viewpoint neutral and therefore subject to intermediate scrutiny); Pittsburgh Press Co. v. Human Rel. Comm'n, 413 U.S. 376 (1973) (holding that ""discrimination in employment is not only commercial activity, it is illegal commercial activity" and newspapers could be prohibited from publishing advertisements for employment that discriminated on the basis of sex).

[179] See Yosemite Park & Curry Co. v. United States, 582 F.2d 552, 558 (Ct. Cl. 1978) ("We begin, as did the Government, with 41 U.S.C. § 252(a), which states unequivocally that executive agencies shall make

Administration's additional work through the White House Office of Management and Budget to develop standards around federal agency procurement and use of AI[180]). The government might also prosecute companies that market their models for illegal purposes or market them in a deceptive or unfair manner, or individuals or entities who use models for illegal purposes. Each of these goals can be pursued with minimal if any constitutional concern.

Recommendations for Minimizing First Amendment Issues with Model Regulation

To the extent the government chooses to target action directly at the publishers or publication of OFMs rather than focusing on context-specific requirements, below is a summary of key factors they should consider in light of the above First Amendment values and precedents. As quickly as the technology is moving, it is possible there may soon be novel facts justifying changes in doctrine — or the Supreme Court may make relevant doctrinal changes in imminent decisions. Therefore this guidance is based on current doctrine, assuming courts find OFMs and model weights to be expressive and protected by the First Amendment. Moreover, these are highly general statements whose application will depend on the particular policy and facts at issue.

- Post-publication regulation or liability is more likely constitutional than a prior restraint such as pre-licensing.

- A prior restraint with strong procedural safeguards including clear criteria, time limits, and opportunity for judicial review, is more likely to be constitutional than one without.

- Restricting commercial speakers is more likely constitutional than restricting non-commercial speakers.

- Requiring transparency or impact assessments around non-controversial objective facts about an OFM's development or performance is more likely to be constitutional than around editorial decisions about what expressive content an OFM can output.

- Similarly, requiring transparency or impact assessments around functional aspects of an OFM is more likely constitutional than around what expressive content an OFM can output.[181]

---

all purchases of goods and services in compliance with the procurement statutes and implementing regulations . . . except where those statutes and regulations are 'made inapplicable pursuant to . . . any other law.'") (second alteration in original)); See also, Regan v. Taxation with Representation, 461 U.S. 540 (1983) (holding that the government need not subsidize all speech).

[180] Executive Office of the President, Office of Management and Budget. "Proposed Memorandum for the Heads of Executive Departments and Agencies," November 2023. https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf. [perma.cc/Y4DN-4GTW]

[181] This is doctrinally distinct from, although potentially factually co-extensive with, the previous consideration.

- Depending on the scope of the requirements, requiring transparency to a regulator for a specified compelling government interest is more likely constitutional than requiring transparency to a broader audience.

- Finally and most importantly, policy interventions narrowly tailored to address evidence-based problems with evidence-based solutions are more likely constitutional than broad interventions aimed at a range of speculative risks.

This last factor of narrow tailoring is likely to be an especially critical question for OFM regulation because at present many proposed policy solutions appear to address speculative risks such as those discussed in part II that are not yet supported by evidence, seek to regulate based on features of OFMs that do not clearly correlate to specific risks (such as the number of floating point operations used to train the model), aim to require evaluations and safety measures that, as discussed above, are still emerging and may not effectively measure for or address the risks intended, and/or seek to prohibit publication of a general-purpose informational asset that could drive expression and innovation in a wide variety of fields and for large numbers of people without a showing that such a broad-brush solution is needed to address known risks. For all of these reasons, courts would likely be skeptical of such expansive and broadly targeted regulatory efforts at this time.

Further considering the above factors, legislators in particular should be careful to draft with severability in mind so that, even if certain requirements with respect to OFMs are found to be unconstitutional, the remainder can stand.

IV. CONCLUSION

NTIA should ensure that its recommendations go through a robust interagency process that includes all of the various agencies with equities in this complex issue, including those with responsibility for competition policy, civil rights, and scientific research — and not just the agencies that oversee national security. Similarly, an opportunity for public comment and a robust interagency process will be vital if the Commerce Department's Bureau of Industry and Security proposes export controls on AI models.

We also urge continued in-depth engagement with civil society on these challenging questions, and appreciate the work NTIA has already done to engage a range of voices. CDT looks forward to continuing to work collaboratively toward AI policies that are evidence-based and effective at protecting the full range of communities impacted by this technology.

***

We appreciate NTIA's solicitation of feedback from stakeholders on these important matters. For additional information, or any inquiries, please contact Kevin Bankston (kbankston@cdt.org), CDT's Senior Advisor on AI Governance.