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

The Center for a New American Security (CNAS) welcomes the opportunity to provide comments in response to NTIA's Request for Public Input on "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights." CNAS is an independent, bipartisan organization dedicated to developing bold, pragmatic, and principled national security solutions. CNAS has several research initiatives focused on critical and emerging technologies, including a center wide, multi-year initiative addressing the national security risks and opportunities of artificial intelligence (AI).

Author: Caleb Withers, Research Assistant, Technology and National Security Program (cwithers@cnas.org)
Thank you to Paul Scharre, Executive Vice President and Director of Studies, for valuable input and feedback.

This document reflects the personal views of the author alone. As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its website annually all donors who contribute.

This response sometimes refers to models with widely available weights as 'downloadable' models.

1.a. Is there evidence or historical examples suggesting that weights of models similar to currently-closed AI systems will, or will not, likely become widely available? If so, what are they?

The most informative data points thus far are lag times between the release of OpenAI's state-of-the-art GPT ("Generative Pre-trained Transformer") language models and the availability of downloadable models offering similar performance:

- GPT-3 was released in June 2020,[1] and Meta's OPT-175B was released in May 2022 for a lag of **23 months**.[2]
- GPT-3.5 was released in March 2022,[3] and Meta's Llama 2 was released in July 2023 for a lag of **16 months**.[4]
- GPT-4 was released in February 2023,[5] **14 months ago. It is yet to be matched** by downloadable models.[6]

Historical evidence prior to the release of GPT-3 is likely less informative. With academia and its norms of open science playing a greater role at the capabilities frontier, weights of state-of-the-art generative AI models were often released publicly.

We can also observe similar time lags following the release of OpenAI's state-of-the-art DALL-E image generation models.

[1] Devin Coldewey, "OpenAI Makes an All-Purpose API for Its Text-Based AI Capabilities," *TechCrunch*, June 11, 2020, https://techcrunch.com/2020/06/11/openai-makes-an-all-purpose-api-for-its-text-based-ai-capabilities/.
[2] Ben Cottier, "The Replication and Emulation of GPT-3," Rethink Priorities, December 21, 2022, https://rethinkpriorities.org/publications/the-replication-and-emulation-of-gpt-3. Cottier makes the case for OPT-175B being the first downloadable model to offer similar performance to GPT-3.
[3] Mohammad Bavarian et al., "New GPT-3 Capabilities: Edit & Insert," OpenAI, March 15, 2022, https://openai.com/blog/gpt-3-edit-insert. These GPT-3 davinci-002 models were subsequently categorized as 'GPT-3.5' models: "Azure OpenAI Service Deprecated Models," Microsoft Learn, February 26, 2024, https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/legacy-models.
[4] "Meta and Microsoft Introduce the Next Generation of Llama," Meta, July 18, 2023, https://about.fb.com/news/2023/07/llama-2/. In identifying Llama 2 as the first downloadable model to offer near-GPT-3.5 level performance, I considered: David Rein et al., "GPQA: A Graduate-Level Google-Proof Q&A Benchmark" (arXiv, November 20, 2023), https://doi.org/10.48550/arXiv.2311.12022; LMSys, "Chatbot Arena Leaderboard," Hugging Face, accessed March 20, 2024, https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard; "MMLU Benchmark (Multi-Task Language Understanding)," Papers with Code, accessed March 21, 2024, https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu.
[5] Tom Warren, "OpenAI Reportedly Warned Microsoft about Bing's Bizarre AI Responses," The Verge, June 13, 2023, https://www.theverge.com/2023/6/13/23759348/openai-microsoft-bing-ai-warning-gpt-4.
[6] LMSys, "Chatbot Arena Leaderboard"; "MMLU Benchmark (Multi-Task Language Understanding)"; Rein et al., "GPQA."

- DALL-E was released in January 2021,[7] and CogView2 was released in May 2022 for a lag of **17 months**.[8]
- DALL-E 2 was released in April 2022,[9] and Stable Diffusion XL was released in July 2023 for a lag of **15 months**.[10]

1.b. Is it possible to generally estimate the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks? How do you expect that timeframe to change? Based on what variables? How do you expect those variables to change in the coming months and years?

To date, well-resourced AI labs—Meta, in particular—have shown a strong commitment to releasing increasingly powerful downloadable models.[11]

Nonetheless, developers of downloadable models will face growing challenges in keeping pace with the AI frontier, potentially increasing the lag time to deploy comparably performing, downloadable models. These challenges include growing costs, secrecy around model algorithms and training, limits to the current data regime, and increased competition.

Growing costs

Increased spending on training has been the largest driver of progress in cutting-edge AI capabilities.[12] If current spending trends continue, the cost for training models will exceed $1 billion within a few years.[13] As a result, several labs may no longer be able to afford training near the AI frontier, especially if releasing their model weights reduces potential monetization.

---

[7] Khari Johnson, "OpenAI Debuts DALL-E for Generating Images from Text," VentureBeat, January 5, 2021, https://venturebeat.com/business/openai-debuts-dall-e-for-generating-images-from-text/.
[8] "MS COCO Benchmark (Text-to-Image Generation)," Papers with Code, accessed March 21, 2024, https://paperswithcode.com/sota/text-to-image-generation-on-coco; Nathan Benaich and Ian Hogarth, "State of AI Report 2022," October 11, 2022, 35, https://www.stateof.ai/2022.
[9] Kyle Wiggers, "OpenAI Expands Access to DALL-E 2, Its Powerful Image-Generating AI System," TechCrunch, July 20, 2022, https://techcrunch.com/2022/07/20/openai-expands-access-to-dall-e-2-its-powerful-image-generating-ai-system/.
[10] Benj Edwards, "Stability AI Releases Stable Diffusion XL, Its Next-Gen Image Synthesis Model," Ars Technica, July 27, 2023, https://arstechnica.com/information-technology/2023/07/stable-diffusion-xl-puts-ai-generated-visual-worlds-at-your-gpus-command/. Pre-XL Stable Diffusion models appear to have performed close to DALL-E 2 (Center for Research on Foundation Models, "Holistic Evaluation of Text-To-Image Models (HEIM)," August 18, 2023, https://crfm.stanford.edu/heim/latest/?group=core_scenarios); given Stable Diffusion XL significantly improved on this models (Dustin Podell et al., "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis" (arXiv, July 4, 2023), https://doi.org/10.48550/arXiv.2307.01952), it seems reasonable to assume it offers at least similar performance to DALL-E 2.
[11] Jonathan Vanian, "Mark Zuckerberg Indicates Meta Is Spending Billions of Dollars on Nvidia AI Chips," CNBC, January 18, 2024, https://www.cnbc.com/2024/01/18/mark-zuckerberg-indicates-meta-is-spending-billions-on-nvidia-ai-chips.html.
[12] "AI Trends," Epoch, accessed March 18, 2024, https://epochai.org/trends.
[13] Paul Scharre, "Future-Proofing Frontier AI Regulation" (Center for a New American Security, March 2024), https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation.

Mistral's recent release of its Mistral Large language model illustrates this dynamic.[14] Before this, Mistral had released the weights of its prior flagship models. If Mistral had done so for Mistral Large, it would likely be among the most powerful downloadable models.[15] Instead, Mistral limited access to Mistral Large through online interfaces and APIs, including through a new partnership with Microsoft. On X (formerly Twitter), Mistral's CEO asked "for a little patience, [as 1,500 NVIDIA H100 GPUs] only got us that far."[16]

Growing secrecy around model algorithms and training

Many key insights underpinning the performance of GPT-3 and subsequent large language models (LLMs) were widely publicized. In particular, GPT-3's architecture was similar to previous models including GPT-2 (2019) and GPT-1 (2018)—just scaled up.[17] These GPT models, along with all of the most powerful LLMs to date, apply the transformer architecture, publicly detailed by Google in a 2017 research paper.[18] Other key insights at Google and OpenAI driving progress in state-of-the-art language models were also detailed publicly, including:

- OpenAI's use of Reinforcement Learning from Human Feedback (RLHF) to train language models to follow instructions.[19]
- Google's 'Chinchilla' scaling laws, which advanced empirical understanding of "optimal model size and number of tokens for training a transformer language model under a given compute budget."[20]

More recently, leading models like GPT-4, Google's Gemini, and Anthropic's Claude 3 have been released *without* detailed discussion of their architecture or training (although there were credible leaks

[14] Romain Dillet, "Mistral AI Releases New Model to Rival GPT-4 and Its Own Chat Assistant," TechCrunch, February 26, 2024, https://techcrunch.com/2024/02/26/mistral-ai-releases-new-model-to-rival-gpt-4-and-its-own-chat-assistant/.
[15] "Au Large," Mistral AI, February 26, 2024, https://mistral.ai/news/mistral-large/; LMSys, "Chatbot Arena Leaderboard."
[16] Arthur Mensch (@arthurmensch), "Clarifying a couple of things since we're reading creative interpretations of our latest announcements: [...]", X (formerly Twitter), February 28, 2024, https://twitter.com/arthurmensch/status/1762818733016322168.
[17] Tom B. Brown et al., "Language Models Are Few-Shot Learners" (arXiv, July 22, 2020), https://doi.org/10.48550/arXiv.2005.14165.
[18] Ashish Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, ed. I. Guyon et al., vol. 30 (Curran Associates, Inc., 2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf; Steven Levy, "8 Google Employees Invented Modern AI. Here's the Inside Story," *Wired*, March 20, 2024, https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper/.
[19] Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback" (arXiv, March 4, 2022), https://doi.org/10.48550/arXiv.2203.02155.
[20] Jordan Hoffmann et al., "Training Compute-Optimal Large Language Models" (arXiv, March 29, 2022), https://doi.org/10.48550/arXiv.2203.15556; nostalgebraist, "Chinchilla's Wild Implications," AI Alignment Forum, July 30, 2022, https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications.

for GPT-4).[21] If this shift to greater secrecy continues in tandem with the value of the underlying intellectual property, it could impede competitors from catching up—especially if leading labs further tighten operational and information security.

Limits of the current data regime

In recent years, labs have trained LLMs primarily on publicly available text, with a particular emphasis on higher-quality, user-generated content (e.g. upvoted posts on Reddit), along with books, scientific papers, code, and other higher-quality websites.[22] For models trained through at least 2022, there was enough higher-quality data available. The bottleneck was scaling up training compute, not data.[23] Competing labs could approximate GPT-3's mix of training data, which drew primarily on publicly available data as OpenAI explained in its release paper.[24]

However, as the amount of compute required to train leading models increases, the availability of higher-quality data has emerged as a constraint.[25] As with model architecture and training more generally, leading labs are no longer detailing their training datasets in public.[26] Additionally, performance in specialized tasks is increasingly driven by training on specialized datasets.[27] Going forward, the most powerful models will employ new training architectures that leverage available datasets more efficiently, or train on novel, non-public, or synthetically generated data.[28] This may prove challenging for some competitors, especially given the greater cost of strategies that rely on purchasing or generating data.

---

[21] OpenAI et al., "GPT-4 Technical Report" (arXiv, March 4, 2024), https://doi.org/10.48550/arXiv.2303.08774; Gemini Team et al., "Gemini: A Family of Highly Capable Multimodal Models" (arXiv, December 18, 2023), https://doi.org/10.48550/arXiv.2312.11805; Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," March 2024, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf; Dylan Patel, "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE," SemiAnalysis, August 28, 2023, https://www.semianalysis.com/p/gpt-4-architecture-infrastructure; "Commoditizing the Petaflop — with George Hotz of the Tiny Corp," Latent Space, March 14, 2024, https://www.latent.space/p/geohot.

[22] Although not necessarily non-copyrighted or distributed legally: Jack Bandy, "Dirty Secrets of BookCorpus, a Key Dataset in Machine Learning," Towards Data Science, May 17, 2021, https://towardsdatascience.com/dirty-secrets-of-bookcorpus-a-key-dataset-in-machine-learning-6ee2927e8650.

[23] Pablo Villalobos et al., "Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning" (arXiv, October 25, 2022), https://doi.org/10.48550/arXiv.2211.04325.

[24] Brown et al., "Language Models Are Few-Shot Learners."

[25] Epoch, "AI Trends"; Patel, "GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE."

[26] See footnote 21.

[27] Baptiste Rozière et al., "Code Llama: Open Foundation Models for Code" (arXiv, January 31, 2024), https://doi.org/10.48550/arXiv.2308.12950; Gemini Team, "Gemini," sec. 4; Zibin Zheng et al., "A Survey of Large Language Models for Code: Evolution, Benchmarking, and Future Trends" (arXiv, January 8, 2024), https://doi.org/10.48550/arXiv.2311.10372; Suriya Gunasekar et al., "Textbooks Are All You Need" (arXiv, October 2, 2023), https://doi.org/10.48550/arXiv.2306.11644.

[28] Villalobos et al., "Will We Run out of Data?"; Anthropic, "The Claude 3 Model Family: Opus, Sonnet, Haiku," sec. 2.5; Dwarkesh Patel, "Will Scaling Work?," April 14, 2022, https://www.dwarkeshpatel.com/p/will-scaling-work.

Increased competition at lower price points

In addition to competing at the frontier, leading labs are increasingly competing on price and speed. In recent months, OpenAI, Google, and Anthropic have released versions of their most powerful AI models that are both cheaper and faster: these models are generally the best available across a wide range of speeds and prices, eroding a traditional competitive advantage of the open-weights ecosystem.[29]

All of these factors may contribute to widening the gap between the capabilities of downloadable models and closed models at the frontier of AI development.

1.c. Should "wide availability" of model weights be defined by level of distribution? If so, at what level of distribution (e.g., 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be "widely available"? If not, how should NTIA define "wide availability?"

Model weights should be considered 'widely available' at levels of distribution even lower than 10,000 entities. Statistically, if we assume that entities with access each have a 1 in 1000 chance of deliberately or inadvertently leaking model weights, the weights will likely be leaked if 693 entities have access. Alternatively, if there is a 1 in 100 chance, a leak is likely if 69 have access.[30]

In practice, distribution of model weights is bimodal:

- Some labs limit external distribution of weights, sharing them only within partnerships with mature Infrastructure-as-a-Service providers (for example, OpenAI's partnership with Microsoft Azure).[31]
- Some labs engage in wider distribution, most often by making them publicly downloadable from the start (e.g. earlier Mistral models, Alibaba Qwen models, or Meta's Llama 2).[32]
    - Meta previously granted researchers and civil society organizations access to some models on a case-by-case basis. The challenges of preventing weight diffusion as more entities are

[29] "Claude 3 Haiku - Quality, Performance & Price Analysis," Artificial Analysis, accessed March 21, 2024, https://artificialanalysis.ai/models/claude-3-haiku#pricing. Note that even where model weights are freely available, there are still hosting and running costs.
[30] $(1-1/1000)^n<0.5$ where $n\geq693$; $(1-1/100)^n<0.5$ where $n\geq69$.
[31] "OpenAI and Microsoft Extend Partnership," OpenAI, January 23, 2023, https://openai.com/blog/openai-and-microsoft-extend-partnership.
[32] Mistral AI, "Open-Weight Models," accessed March 21, 2024, https://docs.mistral.ai/models/; "Qwen," Github, March 21, 2024, https://github.com/QwenLM/Qwen; Meta, "Llama 2," accessed March 21, 2024, https://llama.meta.com/llama2/.

provided access was highlighted when Meta's LLaMA model was leaked on 4chan a week after release.[33]

In addition to the number of entities that have access, NTIA may wish to consider:

- Relevant information security practices and track records of entities with access to model weights.
- The availability of recourse if weights are misused or leaked: e.g. whether measures such as model watermarking are employed to identify leakers, whether there are enforceable legal agreements, and whether entities have adequate means to pay relevant penalties.

2. How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?
The following answer also addresses subquestions 2.a., 2.d., 2.d.i., 2.d.ii. and 2.f.

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

Background on risks from dual-use foundation models

Improvements in foundation model capabilities have outpaced expert expectations.[34] The key drivers of AI progress, including spending on compute and the growing efficiency of chips and algorithms, are all continuing to increase exponentially. If current trends continue, foundation models could be trained on up to a million times more effective compute than today's systems by 2030.[35] Given this pace of progress, it can be simultaneously true that:

- So far, labs releasing foundation model weights have reasonably assessed that the benefits of doing so outweigh the risks.
- Labs are not sufficiently incentivized to account for societal risks in making model release decisions.
- Policymakers should urgently prepare for the possibility of rapid and significant advances in dual-use capabilities of leading foundation models, such that the release of their weights would pose strong societal risks writ large.

[33] "Democratizing Access to Large-Scale Language Models with OPT-175B," Meta, May 3, 2022, https://ai.meta.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/; "Introducing LLaMA: A Foundational, 65-Billion-Parameter Language Model," Meta, February 24, 2023, https://ai.meta.com/blog/large-language-model-llama-meta-ai/; James Vincent, "Meta's Powerful AI Language Model Has Leaked Online — What Happens Now?," The Verge, March 8, 2023, https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.
[34] Ajeya Cotra and Kelsey Piper, "Language Models Surprised Us," Planned Obsolescence, August 29, 2023, https://www.planned-obsolescence.org/language-models-surprised-us/; Katja Grace, "Survey of 2,778 AI Authors: Six Parts in Pictures," AI Impacts (blog), January 4, 2024, https://blog.aiimpacts.org/p/2023-ai-survey-of-2778-six-things.
[35] Scharre, "Future-Proofing Frontier AI Regulation."

Question 2.d. above, along with the 2023 AI Executive Order, appropriately emphasize certain areas where dual-use foundation models could pose particular risk, including:

- Areas where rapid advances in AI capabilities could be especially catastrophic: biotechnology,[36] threats to U.S. national security capabilities, and deceptive misalignment.[37]
- Areas where AI capabilities to analyze data at speed and scale could be transformative, including cyber,[38] surveillance,[39] and military and intelligence operations.[40]

Challenges to mitigating risks when model weights are widely available

Releasing weights for download can make it harder to mitigate certain risks.[41] For instance, when models are accessed through a web interface or API, providers can implement safeguards such as filtering harmful user queries, restricting outputs, monitoring for misuse, and revoking access. But if model weights are downloadable, it is generally straightforward to remove these safeguards.

Most foundation models also undergo specific additional training ('fine-tuning') to reduce their propensity to follow harmful instructions. However, current fine-tuning techniques have largely failed to remove underlying capabilities from the model. Open access to model weights can allow users to reverse safety fine-tuning at relative ease and low cost.[42]

---

[36] S. Alizon et al., "Virulence Evolution and the Trade-off Hypothesis: History, Current State of Affairs and the Future," *Journal of Evolutionary Biology* 22, no. 2 (2009): 245–59, https://doi.org/10.1111/j.1420-9101.2008.01658.x; Christian Ruhl, "Global Catastrophic Biological Risks: A Guide for Philanthropists" (Founders Pledge, October 31, 2023), https://www.founderspledge.com/research/global-catastrophic-biological-risks-a-guide-for-philanthropists; Sarah R. Carter et al., "The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe" (Nuclear Threat Initiative, 2023), https://www.nti.org/wp-content/uploads/2023/10/NTIBIO_AI_FINAL.pdf.

[37] Richard Ngo, Lawrence Chan, and Sören Mindermann, "The Alignment Problem from a Deep Learning Perspective" (arXiv, February 22, 2023), http://arxiv.org/abs/2209.00626; Joe Carlsmith, "Scheming AIs: Will AIs Fake Alignment during Training in Order to Get Power?" (arXiv, November 27, 2023), https://doi.org/10.48550/arXiv.2311.08379; Ajeya Cotra, "Without Specific Countermeasures, the Easiest Path to Transformative AI Likely Leads to AI Takeover," July 18, 2022, https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

[38] National Cyber Security Centre (United Kingdom), "The Near-Term Impact of AI on the Cyber Threat," January 24, 2024, https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

[39] Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence* (W. W. Norton & Company, 2024), sec. III.

[40] Special Competitive Studies Project, "Generative AI: The Future of Innovation Power," September 12, 2023, https://www.scsp.ai/reports/gen-ai/.

[41] Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives" (arXiv, September 29, 2023), sec. 3, https://doi.org/10.48550/arXiv.2311.09227; Kyle Miller, "Open Foundation Models: Implications of Contemporary Artificial Intelligence," Center for Security and Emerging Technology, March 12, 2024, https://cset.georgetown.edu/article/open-foundation-models-implications-of-contemporary-artificial-intelligence/.

[42] Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "LoRA Fine-Tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B" (arXiv, October 31, 2023), https://doi.org/10.48550/arXiv.2310.20624; Xianjun Yang et al., "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models" (arXiv, October 4, 2023), http://arxiv.org/abs/2310.02949; Anjali Gopal et al., "Will Releasing the Weights of Future Large Language Models Grant Widespread Access to Pandemic Agents?" (arXiv, November 1, 2023), https://doi.org/10.48550/arXiv.2310.18233.

As such, labs releasing downloadable models have limited ability to constrain bad actors that may seek to bypass model safeguards. Furthermore, releasing model weights is effectively irreversible—they cannot be recalled if concerning capabilities are discovered or unlocked through new techniques and tools.[43]

Beyond removing safeguards, additional training and fine-tuning can also enhance model capabilities. The U.S. government should regularly assess if adversaries have non-public datasets that could fine-tune leading foundation models in ways that threaten U.S. security. Illustratively, the best coding models have either been, or been derived from, the most capable general-purpose foundation models, which are typically trained on curated datasets of coding data in addition to general training. While sophisticated offensive cyber capabilities have yet to materialize, this stems in part from limited public availability of the most relevant training data, such as exploits and documentation of their development.[44] As such, when downloadable models begin to approach usefulness for sophisticated cyber operations, they may prove more dangerous in the hands of motivated state actors, who can fine-tune them with relevant datasets.

How widely available model weights impact the diffusion of AI capabilities

Training frontier models presents formidable challenges in terms of cost, hardware requirements, data availability, and human expertise. Releasing the weights of these models provides a significant head start to those unwilling or unable to invest the necessary resources to train them from scratch. Where competitors or malign actors leverage these models, there is little that can be done to restrict them from doing so in ways that harm U.S. interests. In its 2023 update to AI chip export controls, the Bureau of Industry and Security specifically highlighted dual-use AI foundation models as examples of the advanced AI systems motivating the new restrictions; widely releasing the weights of models that enable capabilities targeted by these controls risks directly undermining the underlying national security objectives.

When assessing the impacts of releasing model weights, decision-makers will need to account for other factors driving the diffusion of AI capabilities. Insights around model architecture and training techniques can have a more enduring impact than the release of model weights themselves (although releasing a model's weights necessarily reveals its architecture and allows others to experiment with fine-tuning and post-training enhancements)[45]. For example, leading Chinese labs have applied the

[43] Tom Davidson et al., "AI Capabilities Can Be Significantly Improved without Expensive Retraining" (arXiv, December 12, 2023), https://doi.org/10.48550/arXiv.2312.07413; Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety" (arXiv, September 4, 2023), https://doi.org/10.48550/arXiv.2307.03718.
[4444] National Cyber Security Centre, "The Near-Term Impact of AI on the Cyber Threat"; Kumar Shashwat et al., "A Preliminary Study on Using Large Language Models in Software Pentesting" (arXiv, January 30, 2024), https://doi.org/10.48550/arXiv.2401.17459. On coding models and the role of specialized training data for improving specialized tasks, see footnote 27.
[45] See footnote 43.

architecture and training process of Meta's Llama models to train their own models with similar levels of performance.[46]

The ongoing impact of a specific model's weights being available will eventually diminish as the weights of more capable models are released; nonetheless, the pace that this occurs will be influenced by relevant policies.

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

The benefits of foundation models with widely available model weights include:[47]

- Facilitating in-depth research and evaluation: direct, widespread access to model weights themselves allows greater scrutiny of how models' inner workings influence their behavior and capabilities.
- Enabling innovation and customization: developers can build on released model weights to develop new versions and iterations for their specific needs. For example, developers can fine-tune downloadable models on domain-specific datasets, adjusting their behavioral tendencies or response styles, or improving model efficiency while aiming to retain a given level of performance.
- Enabling self-hosting and avoiding lock-in: with the ability to download weights, customers can run models on their own infrastructure, potentially reducing risks from third party access to inputs and outputs, or loss of access following provider outages or model deprecation.

Downloadable model weights are not necessarily a silver bullet for realizing the above benefits. For example, without documentation of training data and processes, users will still be at a disadvantage relative to a model's developers. Moreover, running and fine-tuning the largest foundation models is impracticable on consumer hardware.

[46] Jinze Bai et al., "Qwen Technical Report" (arXiv, September 28, 2023), https://doi.org/10.48550/arXiv.2309.16609; DeepSeek-AI et al., "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism" (arXiv, January 5, 2024), https://doi.org/10.48550/arXiv.2401.02954; 01 AI et al., "Yi: Open Foundation Models by 01.AI" (arXiv, March 7, 2024), https://doi.org/10.48550/arXiv.2403.04652.

[47] Seger et al., "Open-Sourcing Highly Capable Foundation Models," sec. 4; Sayash Kapoor et al., "On the Societal Impact of Open Foundation Models" (arXiv, February 27, 2024), https://doi.org/10.48550/arXiv.2403.07918; Miller, "Open Foundation Models."

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

Widely available model weights do not *differentially* bolster U.S. national security interests: both domestic and foreign researchers, developers, customers, and users can all run, evaluate and build on these models. In some ways, U.S. adversaries are likely to disproportionately benefit:

- It is difficult to meaningfully constrain or monitor users of downloadable models. While many users will have legitimate reasons to prefer this, less restrictive models are nonetheless particularly useful to bad actors and adversaries.
- The U.S. currently leads its adversaries in foundation model capabilities. Countries with weaker models disproportionately benefit from being able to use, build on, and emulate foreign downloadable models.
- U.S. chip export controls constrain China's ability to train compute-intensive foundation models. However, widely available model weights effectively circumvent these controls, allowing Chinese AI labs to download models which they themselves may not be able to train—or for which training may be cost-prohibitive—given these controls.

On the other hand, the availability of U.S. models and architectures makes it tempting for adversaries to rely on them at the expense of their own domestic innovation.[48] As with U.S. chip exports, policymakers must weigh when it makes sense to foster this dependency, when cutting adversaries off may be wise, and to what extent doing so may accelerate indigenous capabilities.[49] As long as the U.S. lead persists, a potentially attractive strategy would be encouraging the diffusion of models and architectures near or slightly ahead of Chinese equivalents, while discouraging this for the most advanced U.S. capabilities.

[48] Kevin Xu (@kevinsxu), "The Beijing Academy of Artificial Intelligence showed Li Qiang this slide during a visit that made it on to CCTV (h/t @niubi) [...]", X (formerly Twitter), March 14, 2024, https://twitter.com/kevinsxu/status/1768365478295355647/photo/1; 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.
[49] Paul Scharre, "Decoupling Wastes U.S. Leverage on China," *Foreign Policy*, January 13, 2023, https://foreignpolicy.com/2023/01/13/china-decoupling-chips-america/; Martijn Rasser, "The Right Time For Chip Export Controls," *Lawfare*, December 13, 2022, https://www.lawfaremedia.org/article/right-time-chip-export-controls.

5.a. What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?

Ideally, model risk evaluations would be *comprehensive*, *interpretable* and *safe*, per the below table.

Comprehensive:	**Interpretable:**	**Safe:**
• Cover as many plausible extreme **threat models** as possible. • Take advantage of **automated and human-assisted** evaluations. • Look at both a model's **behavior** and how it **produced** that behavior. • Use **adversarial testing** to purposefully search for cases where models produce concerning results. • Pursue **robustness against deliberate model deception** to pass evaluations. • **Surface latent capabilities** through practices such as prompt engineering and fine-tuning. • Conduct evaluations **throughout the model lifecycle**. • Study models both **with and without relevant system integrations** (such as external tools or classifiers).	• Include evaluations that present risks in an **accessible** way. • Cover **wide ranges of difficulty** so trends can be tracked over time.	• Ensure evaluations are **safe to implement**: not introducing unacceptable levels of risk themselves.

Desirable qualities of extreme risk evaluations, adapted from Toby Shevlane et al., "Model Evaluation for Extreme Risks" (arXiv, May 24, 2023), http://arxiv.org/abs/2305.15324.

5.b. Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?

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

Methods to remove risky capabilities from trained foundation models have not proved robust, although some recent work has shown tentative promise.[50] Significant further research would be needed before relying on any such method, given the risks involved in mistakenly assuming dangerous capabilities were robustly removed from a downloadable model.

[50] Nathaniel Li et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning" (arXiv, March 6, 2024), https://doi.org/10.48550/arXiv.2403.03218.

5.d. Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?

Almost certainly not. Internet piracy is an instructive case study: while authorities can take legal action in certain circumstances, illegal file-sharing remains rampant, supported by decentralized peer-to-peer file transfer protocols.

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

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

When considering risks from the diffusion of model weights, decision-makers must account for potential unauthorized access or exfiltration. Effective cybersecurity and judicious release policies will both be necessary to address these risks.

Leading foundation models crystallize massive investments in research, engineering, and computation, as well as vast amounts of knowledge, in a few ready-to-run terabytes with potentially thousands of copies—coveted by global adversaries who see AI as a key strategic asset. Where models weights are *not* intended for wide release, securing them from exfiltration will pose a significant challenge, and given the lead times involved, certain efforts may need to be started urgently.[51] Anthropic, for example, acknowledges that it is not yet hardened to a level that would pose significant hurdles to an advanced state actor.[52] Cybersecurity protections must be increased for highly capable models that are not intended to be widely released.

Labs that currently release their leading models for download will have less incentive to invest in security practices preventing model weight exfiltration, all else being equal. If they change their release strategy (or are required to)—perhaps if their models become sufficiently lucrative or powerful—they may face a more formidable challenge in implementing and scaling relevant measures. Standardizing and enforcing cybersecurity practices among leading AI labs now could help establish foundations for stronger future security practices, if needed.

---

[51] Sella Nevo et al., "Securing Artificial Intelligence Model Weights: Interim Report" (RAND Corporation, October 31, 2023), https://www.rand.org/pubs/working_papers/WRA2849-1.html.

[52] Anthropic, "Responsible Scaling Policy," September 19, 2023, https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf. This admission likely reflects *transparency* on Anthropic's part, as opposed to them having worse information security practices than their competitors.

5.f. Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model? To the extent possible, please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.

5.g. Are there means by which to test or verify model weights? What methodology or methodologies exist to audit model weights and/or foundation models?

Greater access to model internals, training details and architectural specifications enables evaluators to conduct more thorough and efficient analyses. Access to internal representations like weights, activations, and attention patterns allows more systematic and efficient explopostration of the range of ways models may respond to different inputs. More generally, understanding a model's training data, architecture, and training process can provide a vital starting point for intuiting how certain capabilities or behaviors may manifest.[53]

Black-box evaluation efforts cannot be expected to successfully surface risks that may be arise when models are later released to millions of users, or that future capability-enhancing methods and tools will unlock.[54] At minimum, the playing field should be even: before any model weights are released, therefore providing users with white-box access and fine-tuning capabilities, evaluators should also be given access to these capabilities, with their findings accounted for in ultimate release decisions.

Of course, this level of evaluator access can itself pose proliferation risks for model weights and details. These risks should be mitigated through technical (e.g. monitored, structured access) and social means (e.g. on-site evaluations, employing a limited pool of trusted evaluators who have signed NDAs).[55] For models weights not proposed to be widely released, more restricted access may be appropriate to guard against proliferation risks.

[53] Stephen Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," arXiv.org, January 25, 2024, https://arxiv.org/abs/2401.14446v1; Benjamin Bucknall and Robert Trager, "Structured Access for Third-Party Research on Frontier AI Models" (Center for the Governance of AI, October 31, 2023), https://www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models; Apollo Research, "We Need a Science of Evals," January 22, 2024, https://www.apolloresearch.ai/blog/we-need-a-science-of-evals.
[54] See footnote 43.
[55] Bucknall and Trager, "Structured Access for Third-Party Research on Frontier AI Models," sec. 5.2; Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," sec. 6.

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

Wide release of model weights offers some of the same benefits of open-source software more generally. However, there are important distinctions in their implications for security.

With traditional software, open-sourcing can help mitigate risks to users from exploitable code: exposing the code to more eyes makes it easier to identify and patch vulnerabilities. In contrast, powerful AI models pose significant risks not just from potential model vulnerabilities, but from the potential misuse of the model's capabilities by operators, including malicious actors. Open-sourcing generally *exacerbates* these risks by increasing the ease of removing or weakening built-in safeguards.

Moreover, whereas typical software features human-written code, the internal representations learned by deep neural networks can be very difficult to explain or interpret—and are certainly not human-readable—reducing the practical benefits of accessing model weights.[56] Unlike traditional, interpretable software logic, deep neural networks' emergent complexity and highly parallel operation make isolating specific 'bugs' or backdoors—let alone formally verifying real-world robustness—impractical.[57] Even if vulnerabilities have been broadly characterized, robustly mitigating them is generally not straightforward.

Where possible, claims about the in principle benefits or risks of downloadable models should be evaluated empirically. For example, white-box access to downloadable models appears to enable the discovery of particularly strong adversarial attacks.[58] Going forward, would-be users and customers will have an interest in knowing how effectively such attacks have been addressed or prevented in a given model, whether downloadable or otherwise.

---

[56] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," ITU Journal: ICT Discoveries 2018, no. 1 (March 1, 2018): 39–48, https://itu-ilibrary.org/science-and-technology/explainable-artificial-intelligence_pub/8129fdff-en; Tim Rudner and Helen Toner, "Key Concepts in AI Safety: Interpretability in Machine Learning" (Center for Security and Emerging Technology, March 2021), https://doi.org/10.51593/20190042; Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models" (arXiv, July 12, 2022), sec. 4.11, https://doi.org/10.48550/arXiv.2108.07258; Elham Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (National Institute of Standards and Technology, January 26, 2023), sec. 3.5, https://doi.org/10.6028/NIST.AI.100-1; Tilman Räuker et al., "Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks," arXiv.org, July 27, 2022, https://arxiv.org/abs/2207.13243v6.

[57] Evan Hubinger et al., "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training" (arXiv, January 11, 2024), https://doi.org/10.48550/arXiv.2401.05566; Apostol Vassilev et al., "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations" (National Institute of Standards and Technology, January 2024), sec. 4, https://doi.org/10.6028/NIST.AI.100-2e2023.

[58] Andy Zou et al., "Universal and Transferable Adversarial Attacks on Aligned Language Models" (arXiv, July 27, 2023), https://doi.org/10.48550/arXiv.2307.15043.

6.c. How, if at all, do licensing arrangements for model weights affect the diffusion of foundation models, or the ability of different entities to benefit from or be harmed by these technologies?

Licensing arrangements will have some risk mitigation benefits, but enforcement will not always be practical, especially for actors that may pose disproportionate risk such as criminal and foreign actors.

7. What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the risks and maximize the benefits of foundation models with widely available weights? What kind of entities should take a leadership role across which features of governance?

Most urgently, the government should prioritize developing robust reporting requirements, expertise, and resources for monitoring and forecasting AI capability trends, assessing key indicators of concern, and formulating appropriate policy responses.[59]

7.a. What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?

Existing executive powers could be used in certain circumstances. For example, export controls on the activities of U.S. persons could apply to the development of models that facilitate WMD-related activities or foreign military, security or intelligence services.[60]

Nonetheless, it would be preferable that Congress provide clear authority, processes, and criteria for when the release of dual-use foundation model weights can and cannot be prevented by the government.

[59] Helen Toner et al., "Skating to Where the Puck is Going: Anticipating and Managing Risks from Frontier AI Systems" (Center for Security and Emerging Technology, October 2023), https://cset.georgetown.edu/publication/skating-to-where-the-puck-is-going/; Anderljung et al., "Frontier AI Regulation," sec. 3.

[60] Emily S. Weinstein and Kevin Wolf, "For Export Controls on AI, Don't Forget the 'Catch-All' Basics," Center for Security and Emerging Technology, July 5, 2023, https://cset.georgetown.edu/article/dont-forget-the-catch-all-basics-ai-export-controls/.

7.d. What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights?

Given the realistic possibility that dual-use foundation models will pose meaningful national security threats in the coming few years, the U.S. government should play a prominent role. Moreover, certain threats cannot be adequately evaluated without government involvement: gauging the extent of AI-enabled cyber threats, for example, will require agencies such as the NSA to be involved.

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

The U.S. has a clear lead in foundation model development,[61] and its chip suppliers dominate the supply of cutting-edge chips.[62] As such, the U.S. should not shy away from measured domestic action to manage risk as it advances the AI capabilities frontier.

Nonetheless, comprehensively grappling with these risks into the future will require multilateral coordination. The United Kingdom and the European Union are natural starting points: like-minded allies with strong AI development and governance capabilities. Beyond this, engagement with China will likely prove necessary.

Additionally, the U.S. should seek shared understanding of relevant risks and benefits with countries occupying key chip supply chain chokepoints—particularly Taiwan, the Netherlands, South Korea, and Japan—to ensure compute governance remains a viable option in the medium-to-long-term.[63]

8. In the face of continually changing technology, and given unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future?

Firstly, evaluation of foundation model is a relatively immature field: interested parties have ample opportunity meaningfully contribute to its rigor and effectiveness.[64]

Secondly, the U.S. government should recognize compute governance as a promising approach for addressing future risks from open foundation models, as it can act upstream of eventual model

[61] LMSys, "Chatbot Arena Leaderboard."
[62] Saif M. Khan, "Securing Semiconductor Supply Chains" (Center for Security and Emerging Technology, January 2021), https://doi.org/10.51593/20190017; Gregory C. Allen, "Choking off China's Access to the Future of AI" (Center for Strategic & International Studies, October 11, 2022), https://www.csis.org/analysis/choking-chinas-access-future-ai.
[63] Girish Sastry et al., "Computing Power and the Governance of Artificial Intelligence" (arXiv, February 13, 2024), http://arxiv.org/abs/2402.08797.
[64] Apollo Research, "We Need a Science of Evals."

development and release. While labs worldwide primarily rely on U.S. chips to train foundation models, the U.S. currently has limited ability to prevent the training or release of models that may harm its interests. The U.S. government (particularly the Bureau of Industry and Security) should pursue a comprehensive, integrated strategy to address risks from exporting AI-related technologies—all the way from semiconductor manufacture through to model weights themselves.[65] This should include:

- Accounting for cloud providers' role in facilitating AI model development and deployment.[66]
- Further action to prevent chip smuggling to U.S. adversaries.[67]
- Advancing the development of secure on-chip governance mechanisms that could support the enforcement of policies like export controls.[68]

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

Firstly, risk management frameworks and regulation should consider the full spectrum of release options between fully closed and fully open. Allowing structured access to highly capable models for developers and researchers through permissive APIs and licenses could realize many benefits of downloadable models, such as facilitating in-depth research, evaluation, and customization (including fine-tuning), while mitigating risks associated with unconstrained public release.[69] In doing so, developers could partner with efforts such as the National Artificial Intelligence Research Resource, reducing dependence on any single company or entity.

Secondly, risk management frameworks and regulations should account for the *marginal* risk that downloadable models pose over existing tools and information, and the extent that relevant defenses can mitigate these risks.[70]

[65] Erich Grunewald and Tim Fist, "Comments on the Advanced Computing/Supercomputing IFR: Export Control Strategy & Enforcement for AI Chips," January 16, 2024, https://downloads.regulations.gov/BIS-2022-0025-0062/attachment_1.pdf.
[66] Lennart Heim et al., "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation" (arXiv, March 13, 2024), https://doi.org/10.48550/arXiv.2403.08501.
[67] Erich Grunewald and Tim Fist, "Comments on the Advanced Computing/Supercomputing IFR: Export Control Strategy & Enforcement for AI Chips," January 16, 2024, https://downloads.regulations.gov/BIS-2022-0025-0062/attachment_1.pdf.
[68] Onni Aarne, Tim Fist, and Caleb Withers, "Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing" (CNAS, January 2024), https://www.cnas.org/publications/reports/secure-governable-chips.
[69]Seger et al., "Open-Sourcing Highly Capable Foundation Models," sec. 4; Bucknall and Trager, "Structured Access for Third-Party Research on Frontier AI Models."
[70] Kapoor et al., "On the Societal Impact of Open Foundation Models."

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

Compute is a useful metric: scaling up of compute has been a key driver of AI capabilities.[71] Even as capabilities are eventually replicated by models with less training compute, new capabilities posing novel risks and regulatory challenges will likely first emerge through large training runs. Additionally, training compute is a relatively straightforward metric, and can be anticipated in advance.

While only a proxy for capabilities, which are what is ultimately of interest, training compute can be a useful filter for which models warrant further evaluation, easing the regulatory burden for the overwhelming majority of AI developers. At the moment, only a small number of AI developers are even capable of training models above the 10^{26} operations threshold.

Models above the compute threshold will need further evaluation with additional metrics, such as directly measuring performance and capabilities, or considering other correlates such as training data or architecture.[72]

Appropriate thresholds will change over time. On the one hand, continued algorithmic progress means that increasingly powerful models will be trainable with a given amount of compute, suggestive of a need to lower thresholds over time.[73] On the other hand, as regulators gather evidence about the societal impacts of increasingly powerful models, as relevant mitigations are able to be implemented, and as increasingly affordable compute ultimately necessitates pragmatism, it may be justifiable to raise thresholds over time.[74] In general, and especially if using more sophisticated metrics, it will be crucial that thresholds are regularly revisited by an appropriately resourced and technically competent regulator.

---

[71] The empirical observation that models that leverage the most computation are the most capable is sometimes known as "The Bitter Lesson", a term popularized by the AI research scientist Rich Sutton: "The Bitter Lesson," Incomplete Ideas, March 13, 2019, http://www.incompleteideas.net/IncIdeas/BitterLesson.html. This observation has been characterized in "scaling laws", which describe how model capabilities scale with training inputs: Villalobos, "Scaling Laws Literature Review," Jan 26 2023, https://epochai.org/blog/scaling-laws-literature-review.

[72] Helen Toner and Tim Fist, "Regulating the AI Frontier: Design Choices and Constraints," Center for Security and Emerging Technology, October 26, 2023, https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/.

[73] Anson Ho et al., "Algorithmic Progress in Language Models," Epoch, March 12, 2024, https://epochai.org/blog/algorithmic-progress-in-language-models.

[74] Konstantin Pilz, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities" (arXiv, February 13, 2024), https://doi.org/10.48550/arXiv.2311.15377; Scharre, "Future-Proofing Frontier AI Regulation," 34.